目录

第1章 素数，因子分解和密码

世界上最大的素数

目前所知的世上最大的素数①硕大无比，用标准的十进位数表示，它有65050位．用指数（或者说幂）的记号写比较简便：

$$2^{216091}-1.$$

即，先用2自乘216091次，然后从中减去1，才得到我们所说的那个素数②．

指数记号容易引起误解．为了对大数的幂表示在脑子里有个较明确的概念，不妨想象一个普通的有8×8个方格的棋盘，并按如下规则往方格里摆放2毫米厚的筹码（如英国10便士的硬币）．先将方格编号，如图1所示从1到64．在第一格里放2枚筹码，在第二格里放4枚，第三格里放8枚．依此类推，下一格里放的筹码数恰为前一格里的两倍．于是，在第n个方格里摞有2^n个筹码．特别在最后一格里摞着2^{64}个筹码．你能想象这摞筹码会有多高吗？1米？100米？1千米？肯定不对！好，不管你信不信，你的这摞筹码将直冲云天，[1] 超过月亮（它只不过400 000千米远），超过太阳（1.5亿千米远），几乎直达（除太阳外）最近的恒星半人马座的α星，离地球大约4光

---

① "素数"这个术语将在下面解释．——原注．

② 这一纪录目前已被三次更新．1996年9月4日，美国威斯康星州克雷研究所的科学家宣布，他们发现了目前已知的最大素数$2^{1257787}-1$，它共有378632位．——译者注．

年．用十进位数表示，2^{64}为

$$18\ 446\ 744\ 073\ 709\ 551\ 616.$$

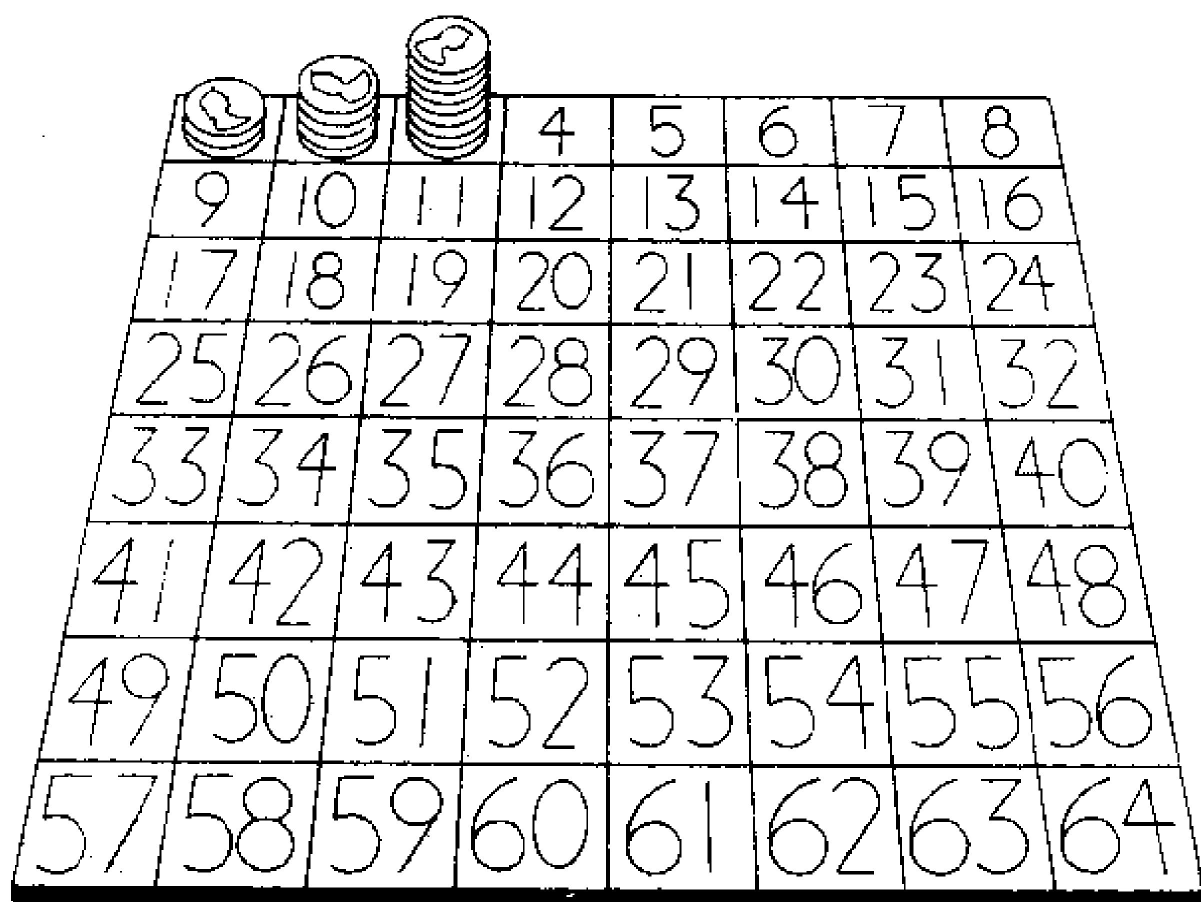


图 1　达到天文数字的棋盘数．第一格里放 2 枚 2 毫米厚的硬币，接下去每一格中摞起的硬币高度都是前一格的两倍，第 64 个方格里摞起的硬币几乎伸展到最近的恒星半人马座的 α 星，大约 4 光年远．

2^{64}就那么可观．为了得到出现在目前最大的素数中的 2^{216091}，你需要一张有 216 091 个方格的棋盘，即在 465×465 个方格的盘上玩上面的把戏！

如何来研究这么大的数呢？首先，你可以使用计算机，但不是随便拿来一台就合用的．上面提到的最大素数是靠世界上最强大的计算机中的一台得到的，这台计算机每秒钟能进行 2000 亿次算术运 [2]

算,它花了3个多小时才完成全部计算.但是单靠机器本身的能力还不够;还需要数学家的技巧.本章的其余部分将讨论这种技巧是如何发展起来的,它还有哪些其他方面的应用.

素 数

"寻找那些最伟大的数是最了不起的行为,从中能获得最大的愉快",弗朗西斯·哈奇森(Francis Hutcheson)在1725年这样写道(《追寻关于善和美的观念的起源》;专论Ⅱ,第3.8节).他心中想的数大概不是数学意义下的当时所知的最大素数,但是他的话用来描述人类对最基本的数学对象——自然数(或者说用于数(shǔ)数(shù)的数)1,2,3,…——的永无止境的迷恋,倒相当贴切.这些抽象的数学对象不仅对我们的日常生活必不可少,而且对几乎所有的数学都是十分重要的,难怪19世纪的数学家L·克罗内克(Kronecker)在谈论数学时说道:"自然数是上帝创造的,其余的一切是人的工作."

适用于自然数的许多性质,都能把它们分成两类(一类具有某个性质,另一类则不具备这个性质).例如,偶性把自然数分成偶数(2,4,6,8,…)和非偶数(即奇数:1,3,5,7,…).再如能被3整除的性质亦然(在本书中凡说到一个数整除另一个数,即指正好除得尽,没有余数.所以,3,6,9,12都能被3整除,而1,2,4,5,7等都不能被3整除).奇-偶的分类既自然又很重要(按能否被3整除所作的分类就不那么自然,也不十分重要).另一种自然而重要的分类是根据完全平方性作出的,像$1=1^2$,$4=2^2$,$9=3^2$,16,25,36,….还有许多其他的分类法,而至今最重要的分类是将自然数分成素数和非素数.

称一个自然数n是素数,如果只有1和n本身能整除它(在这[3]里,1属于特殊情形,习惯上不把1当作素数).

于是,2,3,5,7,11,13,17,19都是素数;1,4,6,8,9,10,12,14,15,16,18,20都不是素数.(有时,我们称那些不是素数的数为合数.)例如,7是素数,因为2,3,4,5,6都不能整除它.

素数之所以重要的主要理由，希腊数学家欧几里得（Euclid，约公元前350～公元前300）就已知晓，他在其《原本》（集中了当时可得到的所有数学知识的13卷本巨著）的第Ⅸ卷中，证明了现在所谓的**算术基本定理**：每一个大于1的自然数，或者是素数，或者可表示为若干素数的乘积，这种表示若不计素数排列的次序则是唯一的.

例如，75900是7个素因子（其中两个各重复出现两次）的乘积：

$$75900 = 2\times2\times3\times5\times5\times11\times23.$$

此式中等号右边的部分称为75900这个数的**素因子分解**.

算术基本定理告诉我们，素数是构作自然数的基本的建材，所有的自然数都是由它们建造的. 素数很像化学家的元素或物理学家的基本粒子. 掌握了任一个数的素因子分解，数学家就获得了有关这个数的几乎全部信息，在本章的后一部分将对此作出明晰的说明（见"密码"一节）. 现在，我们先来了解素数本身.

关于素数，你能问的最基本的问题是它们到底有多少. 比如问有没有最大的素数，或素数是不是无止境的，因而有越来越大的素数？粗粗一看，它们确实非常普遍. 1之后的前10个数（即2至11）中，有5个素数，即2，3，5，7，11，占了总数的一半. 在其后的10个数（12至21）中，有3个素数（13，17，19），占0.3的比例. 在22至31中，素数占的比例仍为0.3. 但在接下去的两组（每组10个数）中，比例降至0.2. 看来，顺着自然数的序列，越往后素数越"稀少". 表1说明，小于n的素数的个数（记作$\pi(n)$），如何随所选的n而变化，并给出了其"密度"$\pi(n)/n$.

n	$\pi(n)$	$\pi(n)/n$
1000	168	0.168
10000	1229	0.123
100000	9592	0.096
1000000	78498	0.078

表1 素数的分布. 对各个n值，给出小于n的素数的个数$\pi(n)$.

可见,顺着数的序列往前走,素数变得稀少了.但是它们最终会[4]完全枯竭吗?回答是否定的.欧几里得已证明了这一结论,他使用的优美的数学推理方法至今仍不失为光辉的典范.首先,想象所有的素数按大小列成表:

$$p_1, p_2, p_3, \cdots$$

所以,$p_1=2$,$p_2=3$,$p_3=5$,等等.目标是证明这张表必定会永远继续下去.换句话说,我们必须证明对任一个 n,已知有素数 $p_1, p_2, p_3, \cdots, p_n$,则在 p_n 之后必然还有不同的素数出现.证明的诀窍是看下面这个数

$$N = p_1 p_2 p_3 \cdots p_n + 1.$$

它是由所有的素数 p_1, p_2, p_3,直至 p_n 的乘积加上 1 得到的.显然 N 比 p_n 大,若 N 恰为素数,则我们知道 p_n 之后确实还有一个素数,这正是我们想证明的.另一方面,如果 N 不是素数,则它必能被某个素数整除,不妨把此素数记作 p.可是你用 $p_1, p_2, \cdots, p_n$ 中的任何一个素数去除 N,皆得余数 1(就是上面我们得出 N 时加的那个 1),因此我们的 p 一定是另一个不同的素数,这样就已经证明了所需的结论.无论如何,总存在一个比 p_n 大的素数,所以我们可以得出结论,素数表将无限延续下去.

注意,我们并不知道上述形式的 N 是不是素数.如果你试算几[5]个例子,你会发现这种形式的数往往是素数.例如,

$$N_1 = 2 + 1 = 3,$$
$$N_2 = 2 \times 3 + 1 = 7,$$
$$N_3 = 2 \times 3 \times 5 + 1 = 31,$$
$$N_4 = 2 \times 3 \times 5 \times 7 + 1 = 211,$$
$$N_5 = 2 \times 3 \times 5 \times 7 \times 11 + 1 = 2311$$

就都是素数.但接下去的 3 个数不是:

$$N_6 = 2 \times 3 \times 5 \times 7 \times 11 \times 13 + 1$$
$$= 30031 = 59 \times 509,$$

$$N_7 = 19 \times 97 \times 277,$$

$$N_8 = 347 \times 27953.$$

事实上，无人知晓是否有无穷多个形如

$$N_n = p_1 p_2 p_3 \cdots p_n + 1$$

的数是素数，同样也不知是否有无穷多个这样的数是合数（当然，这两种尚不确定的情形中必有一种是真实的）. 这只是有关素数的许多陈述简明的问题之一，它们至今都未解决.

涉及素数的最著名的未解决问题之一是哥德巴赫猜想（Goldbach's bonjecture）. 在 1742 年给 L·欧拉（Euler）的信中，C·哥德巴赫提出猜想：每一个大于 2 的偶数皆为两个素数之和. 例如，

$$4 = 2 + 2,$$

$$6 = 3 + 3,$$

$$8 = 3 + 5,$$

$$10 = 5 + 5,$$

$$12 = 5 + 7.$$

[6]

计算机已对 100 000 000 以下的所有偶数作了验算，证明对于这些数哥德巴赫猜想成立；但是时至今日，还没有适当的办法证明整个猜想的正确性.

素 性 检 验

虽然有关素数的大部分经典问题尚未解决，但检验一个数是否素数的方法，在最近几年有了巨大进步. “检验素性的方法？”你一定会大声地问，“那不是很显然的吗？”确实，看一个数是或不是素数，有一种非常自然而直接的方法. 给你一个数，比如说 n，你首先看 2 是否能整除它. 若能，则 n 不是素数，任务完成；若不能，你再用 3 试. 若 3 能整除 n，则 n 不是素数，事情了结；若仍不能，试用 5 去除 n.（你可以跳过 4，因为你已试过 2 不能整除 n，所以 4 也不能整除 n.）若 5 不能整除 n，再试 7.（这回你可以跳过 6，因为 2 和 3 都不能整

除 n.)依此类推,当你试到$\sqrt{n}$而还没找到能整除n的数,那么你就知道 n 必是素数.(因为若 n 不是素数,它将是位于 1 和 n 中间的两个数 u 和v 的乘积,而且 u 和v 中必有一个不大于$\sqrt{n}$.)

上述过程称为试除.这一方法对检验不太大的数据实用,但若数字太大,它就变得十分笨拙.让我们看它会变得多么不合用.假设你在现今最快的计算机上使用高效的程序进行试除(用本章开头提到过的计算机).对于一个有 10 位数字的数,运行该程序几乎瞬间就能显示出结果.对于一个 20 位的数就麻烦一点了,需用 2 个小时.对于 50 位的数则需时一百亿年,令人吃惊.对于 100 位的数,需用的年数为:

1 000 000 000 000 000 000 000 000 000 000 000 000

(这里有 36 个零).这并不只是对于非常大的数作无关紧要的计算.在本章的后面部分将说明,介于 60 位至 100 位数中的素数,是今日密码体系中最安全的一种密码所需要的.[7]

你如何来确定一个 100 位的数是否是素数呢?目前可用的最好方法是 1980 年左右得到的,数学家阿德勒曼(Adleman)、鲁梅利(Rumely)、科恩(Cohen)和伦斯特拉(Lenstra)研究出一种非常复杂的技巧,现在以他们名字的首字母命名为 ARCL 检验.在上面提到过的那类计算机上进行 ARCL 检验,对 20 位的数只消 10 秒钟,对 50 位的数用 15 秒,100 位的数用 40 秒.如果你要它检验一个 1000 位的数,给它一星期的时间也就够了.

这种检验是如何进行的呢?它要依靠相当多的高深的数学——它超出了普通的大学数学的范围,这里不可能给以完全的回答,但解释该方法的中心思想倒不难,它涉及伟大的法国数学家费马(Pierre de Fermat,1601 ~ 1665)的一件简单(而十分精巧)的数学工作.

费马虽是个“业余”数学家(他的专业是法律工作),却得到了数学中一些最精巧的结论,至今仍不失其光辉.他观察到的一个事实是,若 p 是素数,那么对任何小于 p 的数a,数 $a^{p-1}-1$ 能被 p 整除.例如,假定我们取 $p=7$ 和 $a=2$,则

$$a^{p-1}-1=2^6-1=64-1=63,$$

此处 63 确能被 7 整除. 你不妨自己对任意的 p 值(素数)和 a 值(小于 p)试算一下,结论总是对的.

这就给出了一种检验一个数 n 是否素数的可能方法. 先计算 $2^{n-1}-1$,看 n 能否整除它. 若不能,则 n 不会是素数.(因为若 n 是素数,根据费马的观察,你将导出 n 能够整除 $2^{n-1}-1$.)可是,若你发现 n 能整除 $2^{n-1}-1$,又能得出什么结论呢? 很可惜,不能推出 n 必是素数(尽管它好像是素数). 问题出在费马仅告诉我们只要 n 是素数,n 就能整除 $2^{n-1}-1$,而没有说合数都不具有这种性质.(这就好比说所有的汽车都有轮子;这并不排斥其他东西有轮子,比如自行车就是一例.)事实上,确实有非素数具备这种费马性质. 最小的一个是 341,它是 11 和 31 的乘积,所以不是素数;若你(用计算机)去检验一下就会发现 341 能整除 $2^{340}-1$.(过一会儿我们将看到,作此检验并不需要算出 2^{340}的值.)类似于素数而具有这种费马性质的合数,被称为伪素数. 所以,当你用费马的结论检验素性而发现 n 能整除 $2^{n-1}-1$ 时,你能得到的结论是:n 或是素数,或是伪素数.(此时,n 是素数的可能性比较大,因为伪素数虽然也有无穷多个,但出现的频率比真正的素数少得多. 例如在小于 1000 的数中仅有 2 个,小于百万的数中也只有 245 个.) [8]

顺便指出,如果你用其他的数代替 2(比如用 3 或 5)来检验费马性质,不会有什么不同. 无论你用哪个数,都会有伪素数出现,使你在检验素性问题时无法得到绝对肯定的回答.

利用上述检验方法,你无须计算出 2^{n-1};我们已经知道即使 n 不太大,这个数也相当大. 你需要做的只是判断 n 能否整除 $2^{n-1}-1$. 这就是说在计算的任何一步都可以忽略 n 的倍数. 换言之,要计算的是假如用 n 除 $2^{n-1}-1$ 可能出现的余数. 目的是判断此余数是不是零,显然 n 的倍数不会影响这个余数,所以可忽略不计. 数学家(和计算机程序设计者)有标准的符号标记余数:用 B 除 A 所得的余数记作

$$A \bmod B.$$

于是我们就可写出诸如 5mod2 是 1,7mod4 是 3 和 8mod4 = 0.

作为费马检验的一个例子,我们来检验数 61 的素性.我们需要算出数

$$(2^{60}-1)\bmod 61.$$

若此数不为零,则 61 不是素数.若它是零,则 61 或是素数或是伪素数(事实上我们已经知道它是真正的素数).我们将避免计算那个大数 2^{60}.我们先看出 $2^6=64$,因此,$2^6\bmod 61=3$.于是,因为 $2^{30}=(2^6)^5$,我们得到,

[9]

$$\begin{aligned}2^{30}\bmod 61 &= (2^6\bmod 61)^5\bmod 61 = 3^5\bmod 61\\ &= 243\bmod 61 = 60,\end{aligned}$$

于是,

$$\begin{aligned}2^{60}\bmod 61 &= (2^{30})^2\bmod 61 = (2^{30}\bmod 61)^2\bmod 61\\ &= 60^2\bmod 61 = 3600\bmod 61 = 1.\end{aligned}$$

因此,

$$(2^{60}-1)\bmod 61 = 0.$$

此处最后的答案是 0,故结论为:61 或是素数或是伪素数.

你也许想自己试算个例子.那么你不妨先验证

$$2^{10}\bmod 341 = 1,$$

然后利用这个事实证明

$$2^{340}\bmod 341 = 1.$$

这个结果告诉你 341 或是素数或是伪素数(如前面已指出过的,341 事实上是伪素数).

ARCL 检验改进了费马检验,它不再受伪素数的"愚弄".这一改进需用到许多高深的数学.(如果你真的想自己领略一下这方法,可以去看科恩和伦斯特拉写的文章"素性检验和雅可比和"(Primality testing and Jacobi sums),刊于数学研究杂志《计算数学》(Mathematics of Computation)1984 年第 42 卷第 297~330 页.)

梅森（Mersenne）素数

ARCL 检验是目前得到的进行一般目的的素性检验的最快的方法.此处的"一般目的"意指它适用于任意给定的数 n.对于具有特殊结构的数,通常存在可供选用的更快的方法,它们是通过利用数的特殊结构而加快检验速度的.这方面最引人注目的例子是形如 2^n-1 的数.如今称这种形式的数为梅森数,以17世纪法国修道士M·梅森的名字命名.

[10]

梅森在1644年出版的著作《物理数学随感》(Cogitata Physica – Mathematica)的序言中称,对于 $n=2,3,5,7,13,17,19,31,67,127,257$,数 $M_n=2^n-1$ 是素数,而对其他所有小于257的数 n,M_n 是合数.他是如何得到这一结论的呢?无人知晓.但他确实惊人地接近了真理.直到1947年有了台式计算机,人们才能检查他的结论.他只犯了5个错误:M_{67} 和 M_{257} 不是素数,而 M_{61},M_{89} 和 M_{107} 是素数.

梅森数提供了一种找出非常大的素数的漂亮方法.函数 2^n 随 n 的增大快速增长,这保证了梅森数 M_n 很快就变得极大,人们便想到去寻找那些使 M_n 为素数的 n.这类素数称为梅森素数.初等代数知识告诉我们,除非 n 本身是素数,否则 M_n 不会是素数,所以我们只需注意取素数值的 n.不过大多数素数 n 也导致梅森数 M_n 是合数,看来寻找适当的 n 并不容易——尽管前几个数让你觉得并不难,因为

$$M_2=2^2-1=3,$$
$$M_3=2^3-1=7,$$
$$M_5=2^5-1=31,$$
$$M_7=2^7-1=127,$$

都是素数.但是好景不长,

$$M_{11}=2047=23\times 89.$$

接下去是 3 个素数:

$$M_{13}=8191, M_{17}=131071, M_{19}=524287.$$

此后梅森素数变得比较难找了.接下去的 5 个使 M_n 为素数的 n 值是 31,61,89,107,127.

第一次看到上面这些数值的大多数人可能立即得出结论:若 p [11] 本身是梅森素数,则 M_p 也是素数.初看起来它确实不错:3 是梅森素数,M_3 也是;7 是梅森素数,M_7 也是;31 是梅森素数,M_{31}也是;对 127 和 M_{127}情况亦然.可是这种模式到此为止.尽管 8191 是梅森素数(等于 M_{13}),可是 M_{8191}(它有 2466 位数字组成)是合数.这是在 1953 年使用一台早期计算机发现的(参见本章有关完全数的一节).

事实上时至今日,只找到了 30 个梅森素数.上面列出的使 M_n 为素数的 12 个 n 值,在本世纪初就知道了.其后的 5 个($n=521$, 607,1279,2203,2281)都是在 1952 年由 R·鲁滨逊(Raphael Robinson)利用 SWAC 型计算机发现的.1957 年,H·里塞尔(Hans Riesel)利用 BESK 型计算机又发现了一个 $n=3217$.1961 年,A·赫维茨(Alexander Hurwitz)利用一台 IBM7090 型计算机得到了 $n=4253$ 和 4423.1963 年 D·吉利斯(Donald Gillies)在 ILLIAC-Ⅱ型机上找到了 $n=9689$,9941 和 11213.B·塔克曼(Bryant Tuckerman)的 IBM360-91 型机在 1971 年捕捉到了 $n=19937$,下一个发现是在 1978 年,那个创纪录的素数成为报纸的头版新闻.报道称两名 18 岁的高中生L·尼克尔(Laura Nickel)和 C·诺尔(Curt Noll),经过 3 年的努力和 350 小时的计算机计算(使用加州大学海沃德分校的 CYBER174 型计算机),发现了有 6533 位数字的梅森素数 M_{21701}.一年后,诺尔改写了这个纪录,找到了有 6987 位数字的素数 M_{23209}.同年晚些时候,该纪录又被突破,这次是位年轻的程序设计员 D·斯洛文斯基(David Slowinski),他在位于威斯康星州齐普瓦瀑布城的克雷研究所工作.利用极强大的 CRAY-1 型计算机,他找到了有 13395 位数字的素数 M_{44497}.1982 年,同一个人和计算机的配合,证明了 M_{86243}(一个有 25962 位数字的数)是素数.此后,在更强大的 CRAY-XMP 型机上,

斯洛文斯基更上一层楼，得到了有39751位数字的素数 M_{132049}. 最后，于1985年9月，在得克萨斯州的休斯敦市，谢夫隆地球科学研究所的CRAY－XMP型机发现了有65050位数字的 M_{216091}，成为当今的最新纪录.（因为谢夫隆研究所使用的是斯洛文斯基设计的“素数发现者”程序，这一发现理应归功于他. 他们之所以愿意用这个程序，原因是它能以有效的方式显示该计算机系统的任何故障.）[①]

故事就到此结束了吗？也许不然. 人们猜想梅森素数是无止境的——它们有无限多个，不过目前尚未证明. 我们确凿知道的是它们至少有30个（即至今已验证过的那些.）

检验梅森数的素性的方法非常简单（其数学背景则不然），现称卢卡斯－莱默（Lucas－Lehmer）检验；E·卢卡斯于1876年发现了其基本思想，D·莱默则于1930年给出了具体方法. 为检验梅森数 M_n 是否是素数（假定已知 n 为素数），可按下述规则计算 $U(0), U(1), \cdots, U(n-2)$ 这些数： [12]

$$U(0)=4,$$
$$U(k+1)=[U(k)^2-2]\bmod M_n.$$

若你最后得到 $U(n-2)=0$，则 M_n 是素数. 若 $U(n-2)\neq 0$，则 M_n 不是素数.

例如，我们要想用卢卡斯－莱默检验判断 M_5 是否素数（因 $M_5=2^5-1=31$，非常简单，我们已经知道它是素数，这里只为讲清楚方法），让我们作下述计算：

$$U(0)=4,$$
$$U(1)=(4^2-2)\bmod 31=14\bmod 31=14,$$
$$U(2)=(14^2-2)\bmod 31=194\bmod 31=8,$$
$$U(3)=(8^2-2)\bmod 31=62\bmod 31=0.$$

① 1992年3月，发现了迄今已知的第31个梅森素数 M_{756839}，它共有227832位；1994年又发现了第32个梅森素数，它有258716位；1996年9月，美国威斯康星州克雷研究所发现了最新的梅森素数 $M_{1257787}$，它共有378632位.——译者注.

由于 $U(3)=0$,M_5 必为素数.

你不妨自己对下面两个数进行检验:$M_7=127$(这是素数)和 $M_{11}=2047$(这不是素数,参见前面).

因子分解

在颇具声望的美国数学会于1903年10月举行的会议日程上,列有数学家F·N·科尔(Frederick Nelson Cole)提交的一篇文章,它的题目相当平凡:“关于大数的因子分解”.当轮到他讲演时,科尔走到黑板前,一言未发便作起2的幂次的演算,直至2的67次幂,然后从所得结果中减去1;他沉默无言地走到黑板的空白处,又将下述两数相乘:[13]

193707721 和 761838257287.

两次演算的答案是一样的.然后他还是未吐一字回到了座位上,这时全场听众站了起来为他热烈鼓掌,这在美国数学会开会的历史上是绝无仅有的一次.

科尔所做的(用了他20年来的所有周日下午)就是寻找梅森数 M_{67} 的素因子.1867年以来,人们已经知道 M_{67} 是合数,那是E·卢卡斯本人用卢卡斯(现称卢卡斯–莱默)检验发现的.该检验虽然能回答给定的梅森数是素数还是合数的问题,但对于合数的因子却一无所知.(ARCL检验也是如此,前面已大致描述了对它的评价.事实上,目前使用的任何快速素性检验法都如此.)

那么,你如何去找合数的因子呢?尝试法显然不在考虑之列,理由跟在讨论素性检验时说明的一样,它不实用.但实际上,目前所有用于素性检验和因子分解的办法中又都少不了试除法.一般在开始阶段它运算起来并不慢,比如先用前一百万个素数做试除.如果找到一个因子,那么素性和因子分解问题就都解决了.若未找到,这时你至少知道这个数或是素数或是合数;而若是合数,它只有极大的素因子.后者是费马考虑的一种简单的因子分解法的出发点,下面就描述

这种方法.

设 $n = uv$，其中 u 和 v 都是大的奇数，并不妨令 $u \leqslant v$.（我们假设 n 只有大的素因子，面临的问题恰好是 n 为合数而要找出它的因子.）令

$$x = \frac{1}{2}(u+v), \quad y = \frac{1}{2}(u-v).$$

于是，$0 \leqslant y < x \leqslant n$，且 $u = x + y, v = x - y$，故

$$n = (x+y)(x-y) = x^2 - y^2,$$

它可改写为

$$y^2 = x^2 - n. \tag{1}$$

反之，若 x 和 y 满足方程(1)，则 n 可分解因子为

$$n = (x+y)(x-y). \tag{2}$$

因此，将 n 分解为两个数的乘积等价于去找出满足方程(1)的数 x 和 y，此时最后的因子分解由方程(2)给出.（注意，这不一定能得到 n 的素因子分解. 不过一旦一个数被分解为两个因子，这两个因子又可能继续分解，此时任务就简单多了，因为数越小分解越容易.）

为了找出满足方程(1)的 x 和 y，可先从 k 开始演算——k 是大于等于$\sqrt{n}$的所有数中的最小者，依次检验 $x = k, x = k+1, x = k+2, \cdots$ 这些数是否能使 $x^2 - n$ 成为完全平方数. 一旦找到这样的 x，当然分解工作就完成了. 倘若 n 有两个大小相近的因子（因此很接近开始时考虑过的$\sqrt{n}$），其中之一应能很快地被找到.（如果你想亲自试试，不妨对 10379 和 93343 试算一下.）

我们有各种办法能加速上述运算过程. 例如，假如你用手算，就没有必要每次都算出 $x^2 - n$ 的平方根以判断其值是不是整数. 因为没有一个完全平方数的最后一位是数字 2，3，7 或 8，所以只要发现 $x^2 - n$ 的末位数字是这几个数，相应的 x 值可立即不再予以考虑.

费马本人利用这一方法得到了如下分解：

$$2027651281 = 44021 \times 46061.$$

计算机使用某些相当复杂的方法可"瞬时排除"不可能的 x 值(这一过程被形象地称为筛法).1974年,加州大学伯克利分校的数学家建[15]造了一台专为筛数用的电子装置 SRS-181,它每秒能筛2千万个数.

费 马 数

第 n 个费马数是这样得到的:它是 2 的某次幂再加 1,该幂次本身是 2 的 n 次幂,即

$$F_n = 2^{2^n} + 1.$$

于是,$F_0 = 3$,$F_1 = 5$,$F_2 = 17$,$F_3 = 257$(由于在重复地使用指数函数,其快速增长的势头已经显露出来),而 $F_4 = 2^{16} + 1 = 65537$.

由于费马在 1640 年给梅森的一封信里提到了这些数,遂引起了人们对它们的兴趣.费马注意到 F_0 到 F_4 这几个数都是素数,他在信中写道:"我已发现形如 $2^{2^n}+1$ 的数永远是素数;很久以前我就向分析学家们指出了这个定理是正确的."这段话对所有只根据少量信息就下结论的人而言,是值得引以为戒的警钟.费马尽管对于数有极强的判断力,但上述说法却是错误的.伟大的瑞士数学家 L·欧拉在 1732 年首先推翻了他的结论:$F_5 = 4294967297$ 不是素数.欧拉是用试除法得到这个结果的,但具讽刺意味的是,用费马自己的检验法直接计算也能证明 F_5 并非素数.你还记得这种检验吧:若 p 是素数,则 $3^{p-1} \bmod p = 1$;对于 $p = F_5$,$3^{p-1} \bmod p = 3029026160$,所以 F_5 不可能是素数.

续后的工作已表明费马在这个问题上是大错特错的.现知对从 5 至 16 的所有 n 值,F_n 皆为合数,对其他许多 n 值亦然;目前猜测对所有大于 4 的 n 值,F_n 全是合数.

费马数还成了另一个实例,其特殊的形式可能有助于有效地来检验它们的素性——这种被普遍接受的方法称为普罗斯(Proth)定理:费马数 F_n 是素数,当且仅当

$$3^{(F_n-1)/2} \bmod F_n = -1.$$

这个结论能非常有效地检验一个费马数的素性.(你可能猜到了,它跟上面讨论过的费马检验关系密切.)但我们现在的兴趣不在检验费[16]马数的素性,而在将已知为合数的数进行因子分解.近年来这一领域有了一些长足的进步,而且在数学王国之外有应用.(参见本章有关密码的一节.)

我们已指出,欧拉证明了费马数 F_5 是合数.欧拉还算出了它的一个素因子是 641.1880 年,兰德里(Landry)证明 F_6 是合数,也找到了它的一个素因子:274177.至于 F_7,情况稍有不同.莫尔黑德(Morehead)和韦斯顿(Western)在 1905 年就证明了它是合数,可迟至 1971 年才由布里尔哈特(Brillhart)和莫里森(Morrison)利用 IBM360-91 型计算机找到了它的因子

$$\begin{aligned} F_7 &= 2^{128}+1 \\ &= 340282366920938463463374607431768211457 \\ &= 59649589127497217 \times 5704689200685129054721. \end{aligned}$$

他们使用了很早由莱默(Lehmer)和鲍尔斯(Powers)提出的一种涉及连分数的方法,计算持续了约一个半小时.连分数法(按照现在的称呼)的改进型提供了目前最好的几种因子分解方法.

就是 1905 年证明 F_7 为合数的莫尔黑德和韦斯顿,又在 1909 年发现 F_8 是合数.可是要到 1981 年才由布伦特(Brent)和波拉德(Pollard)找出其因子分解,整个计算在 Univac1100/42 型计算机上进行了 2 小时.波拉德自己设计的方法不同于大多数数学方法,它不能保证一定算出肯定的结果,依据其数学背景所能得出的结论只是:当执行运算时,极有可能在合理的时间内得到一个数的因子分解,不可能的概率很小.所以,这种方法不同于试除法,后者在 10 亿年内得到解答的机会很小;波拉德的方法虽然也含有机会的因素,但其巧妙之处在于机会偏好于给出可能的因子.近年来出现了一些所谓的蒙特卡洛方法(Monte Carlo method),诸如波拉德的因子分解技术,它们舍去结论成立的完全确定性,换来的是在相当短的时间内得到有极大的

[17] 可能性成立的结论.

F_8(它由 78 位数字构成)的两个素因子是

1238926361552897

和

93461639715357977769163558199606896584
051237541638188580280321.

在写作本书时,尚无人能分解 F_9. 如果德国数学家 K·F·高斯(Karl Friedrich Gauss)能活到现在有了高速计算机的时代,也许对此会有帮助.高斯确实导出过与经典的希腊几何问题相联系的、有关费马数的最令人惊讶的成果.这项成果以及高斯的许多发现,促使我要对这位具有世上最惊人的数学头脑的人作专门的介绍.

惊人的数学头脑

卡尔·弗里德里希·高斯 1777 年生在现位于德国的不伦瑞克市.他的父亲是名瓦匠,希望儿子成为他的帮手,又能干活又能算账.年轻的高斯看来很适合记账,在 3 岁时就曾改正过父亲计算工资时的错误.数学真是幸运(更不必说物理和天文学了),掌权的公爵很快听说了这个孩子的天才,便安排他接受正规教育.15 岁那年,高斯的能力已大大超出了他的老师,于是进了卡罗琳学院.在三年的学院生活期间,那里的教授也不得不承认高斯超过了他们.

1796 年,高斯还是个大学生,他对涉及希腊几何和费马数的一个问题作了令人注目的研究,其成果写进了他的巨著《算术研究》的第 7 节和最后一节.这本书于 1801 年出版,那年高斯仅 24 岁;此书[18] 成为今日数论之基础(数论是数学中研究自然数性质的一个分支学科,本章的内容只涉及其一小部分).

古希腊数学家最喜爱的问题之一是使用直尺(不带刻度,只适合画直线)和圆规(只能用来画圆的弧,不能用于在纸上移动而量出等长线段)画出平面图形(圆,三角形,平行四边形等).通常靠着相当巧

妙的设计，人们仅用那两种最初等的工具就能进行许多几何图形的作图.（直到本世纪60年代中期以前，这类作图是全世界的学生的数学课中重要的内容.）希腊人已经知道如何去画正 n 边形，此处 $n=3,4,5,6,8,10,12,15,16$.（一个多边形称为正多边形，是指它所有的边都等长，所有的内角都相等.）

19岁时高斯证明了：一个正 n 边形可仅用直尺和圆规作图，当且仅当或者 $n=2^k$（k 是某个数）或者 $n=2^k p_1 p_2 \cdots p_r$（k 是某个数），其中 $p_1, p_2, \cdots, p_r$ 是不同的费马素数.特别地，对任一个费马素数 p，你能画出正 p 边形.对于第一个费马素数 $F_0=3$，你能画出等边三角形，这很容易做到；对下一个 $F_1=5$，你可画出正五边形.因为 $F_2=17$ 也是费马素数，所以高斯的结论说明正17边形是可以用直尺和圆规作图的.这是自希腊时代以来在正多边形作图问题上的第一个（也是唯一的）进展.高斯对此发现十分自豪，曾请求在他的墓碑上刻上一个正17边形.这一要求虽未实现，但在他家乡不伦瑞克为他竖立的纪念碑上雕了这样的多边形.

完 全 数

毕达哥拉斯的信徒（公元前6世纪的数学家毕达哥拉斯(Pythagoras)的追随者）注意到，数6有一个特性，它等于它自己的因子（不包括它自身）的和：

$$6=1+2+3.$$

[19]

6之后具有同样性质的下一个数是28.能整除28的数是1,2,4,7,14和28本身，而

$$28=1+2+4+7+14.$$

毕达哥拉斯的信徒称这类数为完全数.

在公元1世纪的一本书《算术入门》中，希腊数学家尼可马修斯(Nicomachus)列出了4个完全数，（6和28之后的）第三个是496，下一个是8128.由此可得出两个猜测：第 n 个完全数含有 n 个数字，完

全数的最后一位数字以6和8交替出现.这两个猜测都是错的.首先,不存在5位数的完全数.进而,虽然第5个完全数确以6结尾,等于33550336;可第6个亦然,等于8589869056.(然而,任何完全数确实都以6或8结尾.这可以直接证明而无需知道哪些具体的数是完全数.)

欧几里得在大约公元前350~300年间,在他的《原本》第Ⅸ卷中证明了:若2^n-1是素数,则数$2^{n-1}(2^n-1)$是完全数.两千年后,欧拉证明每个偶完全数都具有这种形式.这就在梅森素数和完全数之间建立了紧密的联系,它立即能准确地说出存在30个偶完全数.事实上,我们不知道有奇完全数,人们猜想所有的完全数都必定是偶数.虽然这一结论尚未证明,而能收集到的证据全都有利于这个猜想.现知若果然存在奇完全数的话,它必大于10^{100},并且至少有11个不同的素因子.另一方面,历史告诫我们,对完全数作猜想应十分谨慎.P·巴洛(Peter Barlow)在他1811年的著作《数论》中写道,欧拉于1772年发现的有19位数字组成的第8个完全数$2^{30}(2^{31}-1)$"是所能找到的最大的完全数;因为它们只能满足好奇心而没有任何用处,所以,大概不会再有人试图去寻找超过它的完全数了."

巴洛说的完全数只有满足好奇心的价值似乎不错,可是他显然低估了好奇心的作用,本章第1节已充分表明了好奇心的力量.无疑,完全数非常奇特.例如,每一个(偶)完全数皆为三角形数(triangular);即它可由排列成等边三角形状的球的总数来表示(等价于写成$\frac{1}{2}n(n+1)$的形式).另一个事实是:对除6以外的任一完全数,你把它的各位数字相加,其结果必等于9的某个倍数再加1.与此相关的结论是任一完全数的数根(digital root)等于1.(所谓数根,是指你把该完全数的每一位数字相加,再把所得和数的每一位数字相加,直至最后得到的和数为一位数,此一位数即原数的数根.)

[20]

此外,每个完全数可表为连续的奇数的立方和.例如

$$28 = 1^3 + 3^3$$

$$496 = 1^3 + 3^3 + 5^3 + 7^3.$$

还有，若 n 是完全数，则 n 的所有因子的倒数的和永远等于2.例如，6的因子为1,2,3,6,则

$$\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{6} = 2.$$

事实上，尽管巴洛宣称完全数无用，可是人们在花费如此多的精力寻找这些“奇特”的数的时候，其中所作的计算本身已为判断计算机的能力提供了一种标准.拿梅森数 M_{8191} 为例，它是第一个打破使梅森数成为素数的梅森素数链的数(参见本章第4节).为了论证这个有2466位数字的数不是素数(因而也不能产生出一个完全数)，1953年有人首先在ILLIAC－Ⅰ型机上利用卢卡斯－莱默检验进行演算，居然用了100小时.后来计算时间显著缩短了：IBM7090型机用了5.2小时，ILLIAC－Ⅱ型机用了49分钟，IBM360－91型机用了3.1分钟，而CRAY－1型机仅用了10秒钟.

密　码

1982年秋，在加拿大温尼伯市举行的科学会议期间，两位数学家和一位计算机工程师于某天晚上外出喝啤酒.两位数学家很快转入了如何分解大数的因子的话题，自然碰上了计算问题.计算机工程师听着他们的谈话，指出他设计的一台特殊的计算机可能很容易地克服他们遇到的一个主要的困难.这次在啤酒店碰巧发生的事，对数据安全的研究将产生重大的影响.大数的因子分解十分困难，这正是一种最安全的密码体系的关键所在.纯粹数学中一种看来无用而深奥的研究课题，居然成为现代安全体系的基础，这是本世纪数学中发生的最有趣的故事，它也向那些随意宣称某件科学工作“毫无实用价值”的人敲响了警钟.

〔21〕

某些最积极地诋毁数学的效用的人恰好是数学家．伟大的英国

数学家哈代（G. H. Hardy）在他著名的小册子《一个数学家的辩白》（*A Mathematician's Apology*）中写道："真正的数学对战争没有影响．还没有人发现数论或是相对论服务于战争目的，在许多年内似乎也不会有人发现这种事"(见该书第 28 章)．这段话写于1940年．到 1945 年，世界已目睹了哈代关于相对论对战争无用的可怕的否证：原子弹爆炸了．至于他举的另一个例子——数论，这门"无用"的学科所提供的各种安全体系，正用于控制（也许某一天用于发射）成百颗核子导弹；自从在广岛投下第一颗原子弹后，核导弹的数目已大大增加了．数学的发现在整个世界中到处都有可以预见的（或所需的）应用．碰巧，哈代本人从事的正是数论研究，他自己的某些工作已被证明有实用价值，尽管他个人宣称（见同一小册子的第 29 章)："我从未做过任何'有实用价值'的事情．没有一项我的发现，对世界的舒适程度产生过（或可能产生）哪怕是最小的，直接或间接的，好的或坏的影响．"

当然，编制密码的思想由来已久．儒略·恺撒(Julius Caesar)使用密码是为了在高卢战争期间安全地传送命令给他的将军．今天，不仅军方需要用加密技术保证通讯的安全，商界和政界也要求保证信息不落入他人之手．

你如何来设计编码(即加密)体系呢？光回答"多加小心"是不够的．可能的密码分析人员(即试图破译你的码的"敌人")有大量的武器可供其使用，即有强大的计算设备，又有复杂的数学和统计技术．[22] 恺撒使用过的那类简单的密码肯定极不安全．恺撒暗码是这样编制的：原信息中的每个词中的字母，按照某个固定的规则，依次用另外的字母代替，比如用位于其后三位的字母表中的字母代替．于是用 D 代替 A，用 J 代替 G，用 B 代替 Y，等等．MATHEMATICS 这个词就变成了 PDWKHPDWLFV．表面上看，若你不知道所用的法则，按此法加密的信息就完全无法解密，但情形并非如此．一方面，这种"替换"码只有 25 种，怀疑到你使用了这类码的敌人，只需要一种一种地试就能找到你使用的那种替代规则．即使你使用不那么简单的字母替代

规则，所编制的码仍不安全. 问题出在英语中单个字母以非常确定的频率出现（在其他语言中也有同样的现象），只要计算每个字母在码文中出现的次数，敌人很容易推断你所使用的替换规则，特别当使用计算机加速这种推断过程的时候.

简单的字母替代法被排除之后，你还能试用什么办法呢？不管你选择哪种方法，同样的危险仍会出现. 只要在你编好的码文中有某种“可被辨认”的模式存在，高级的统计分析方法一般不难破译你的密码. 现在，真正的困难之所在变得明朗了. 为了使你的信息能在接收者那里（可能在数千英里之外）被正确地解码，你在对信息用你的加密方案进行变化时，显然不能排除所有的规则——信息必然要受到某种规则的支配. 这种隐藏着的规则应埋得足够深，以防被敌人发现.

所有现代的密码体系都要使用计算机；它们必须如此. 一般都假定敌人拥有强大的计算机来分析你的信息，所以你的体系必须足够复杂，以防计算机的攻击. 为了使所设计的密码体系尽可能安全，它们必定由两部分构成：一个加密程序和一把“钥匙”. 前者是典型的计算机程序，也可能是一台专门设计的计算机. 为了给信息加密，该体系不仅需要这些程序，还要一把选择好的钥匙，它通常是一个秘密选定的数. 加密程序将依赖这把选定的钥匙对信息编码，使得只有知道这把钥匙的人才可能解开所编的码文（参见图 2）. 由于安全性依赖于这把钥匙，所以可以有许多人在一段相当长的时期里，使用同一个加密程序，这就意味着值得花大量的时间和精力来设计这种程序. 不妨看一个有助于理解的类比. 生产保险柜和锁的工厂可以设计一种 [23] 类型的锁卖给几百个使用者，后者靠自己独特的钥匙来保证安全.（此处说的“钥匙”可以是用于字码锁的各种字码，于是立即显示出“钥匙”这个词的两种用法间的类似.）正如敌人可能知道你的锁是如何设计的，可是不知道锁的字码仍无法打开你的保险柜，所以，敌人可能知道你所用的加密体系而无法破译你的编码信息——要想破译就得知道你的钥匙.

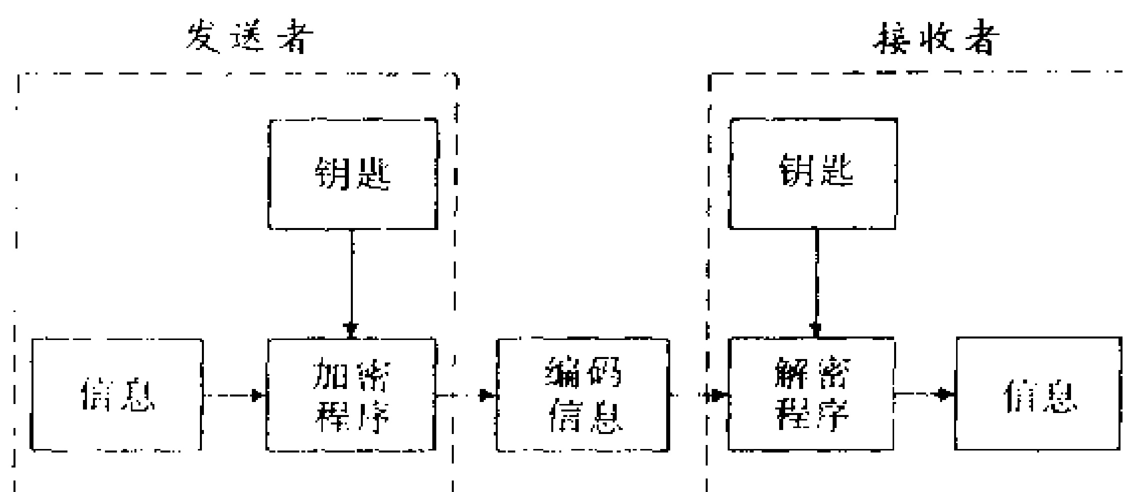


图2 典型的密码体系. 加密程序(它可以是一种特别制造的装备,或是可在通用计算机上运行的程序)要利用使用者选定的密钥以编制码文. 在接收端,有一类似的体系运作. 传统的体系在加密和解密时使用同样的钥匙,解密程序只是做加密程序所做的相反的工作. 公开密钥体系则使用两种不同的钥匙,编码和解码间的关系取决于所利用的数学.

在典型的"密钥体系"中,信息发送者和接收者事先要协商好某种密钥,然后利用它互送信息. 只要他们保守钥匙的秘密,该体系应该(如果它设计得很好)是安全的. 美国人设计的数据加密标准(DES)就属于这种体系,它的钥匙是个数,其二进位制表示有56位(换言之,这把钥匙是由56个0和1组成的数链). 为什么要这么长的钥匙? 好,现在来解释. 实际上没有人对DES体系是如何工作的加以保密,一切细节都是公开的. 从理论上讲,任何敌人只要试遍所有可能的钥匙就能找到哪把钥匙在起作用. 就DES而言,共有2^{56}种可能的钥匙,这个数如此之大,以至想要试遍所有的钥匙实际是不可能的.(事实上这个数字还没大到足以提供绝对的安全,不过对任何密码体系都必须兼顾安全性和使用者方便两个因素. 钥匙越长,使用就越不方便.)

[24]

虽然目前广泛使用着DES,但该体系有着明显的欠缺. 在使用前,发送者和接收者必须协商好他们将使用的密钥;因为不愿通过任

何通讯渠道传送密钥，他们必须碰面并选定钥匙，或者起码要雇用一位可信任的信使来传送密钥．所以，这种体系不适合未曾谋面的个人之间的通讯．特别地，它不适合诸如国际间的银行及商务活动，而他们往往需要发送保密信息给世界各地的发送者从未见过面的人．

1975年，W·迪菲（Whitfield Diffie）和M·赫尔曼（Martin Hell－man）提出一种新型的密码体系：公开密钥的密码学，其中的编码方法需要两种钥匙而不只是一种：其一用于加密，其二用于解密（就好像一把锁要用一把钥匙把它锁上，用另一把钥匙把它打开）．这种体系的使用方法如下：新的使用者要购买一本供该通讯网络的所有成员使用的标准程序（或专用计算机）．然后他应确定两把钥匙．一把是他的加密钥匙，他应严加保密．另一把用于破译网络的任一使用者发送给他的编码信息，这把钥匙将刊出在网络使用者手册上．为了发送一个信息给网络使用者，需要做的全部工作就是查出那位使用者所公开的那把密钥，用它对信息加密，并发送出去．对任何人而言，知道这把公开的密钥对破译密码是毫无帮助的．解码需要另一把专门解码用的钥匙，而这只有那位接收者知道．（所以信息一旦被加了密，那位信息发送者也无法破译它！）

好，看来这办法很不错，但是如何具体地实现这样的体系呢？乍看起来，这似乎是不可能的：所用的钥匙（恕我在这里用这个词）要同时利用令编码者的任务变得如此艰巨的那些计算机的功能和不足．本章前面的部分指出过，找出大的素数（比如50位数字大小的）相对而言比较容易．把如此大小的两个大素数相乘得出一个（合）数（大约有100位或更大些）也不难．但要把这么大的数分解成两个素因子就难上加难了，无论出于什么意图和目的，事实上都是不可能做到的．这就是当今普遍使用的公开密钥体系所依赖的基本思想，这一体系由麻省理工学院的R·里弗斯特（Ronald Rivest），A·沙米尔（Adi Shamir）和L·阿德勒曼（Leonard Adleman）设计，现以他们姓氏的首字母命名为RSA体系．需要保密的用于解密的钥匙（本质上）由两个大的素数组成．（使用者要借助于计算机选出它们，而不能在任何公开

25

出版的素数表中去选.后者敌人很容易弄到手!)公开的用于加密的钥匙就是这两个素数的乘积.因为不存在快速的大数因子分解的方法,所以实际上不可能根据公开的加密钥匙重新找到解密钥匙.信息的加密对应于两个大素数相乘(容易),解密对应于相反的因子分解过程(困难).(这还不是此类体系的全部工作原理.我们还需要某些数学知识,它们虽然比较复杂,但费马就已经都了解了.重要之点在于,因为解密过程正好跟加密过程相反,所以两种钥匙间的关系也必然如此.)

这就是当今庞大的国际数据通讯网络能安全运行的原因,它依赖的是数学家的一种无能:他们尚未找到大数因子分解的有效方法(同时却能容易地找到大的素数).显然,这类体系的安全(除此处未提及的外)取决于不断出现的因子分解方面的困难.温尼伯啤酒店的对话就由此而起.最初,体系的设计者建议用两个各约有50位数字的素数,这已经能提供足够的安全性(对这类系统,使用的数越大,运作起来越昂贵,所以要寻找一种理想的折中方案).直到1982年,最好的因子分解方法只能处理约有50位数字的数(使用诸如CRAY-1型计算机).计算机工程师T·沃诺克(Tony Warnock)告诉因子分解专家M·文德利希(Marvin Wunderlich)和G·西蒙斯(Gus Simmons)的信息是,特殊设计的CRAY-1的运算器可以解决他们的一个问题,即能使他们的方法分解有60位至70位数字的大数.这样,RSA体系的安全就受到了威胁.虽然对付的办法很显然——只要使用每个都有100位数字的素数,以得到一个有200位数字的公开密钥;但是,这次事先未料到的交谈确实在安全通讯的事业中搅起了一片易变的涟漪.当前在因子分解方面的其他进展也无助于克服这种"不安全感".虽说目前一般认为因子分解的上限是有90位数字的数,可是大量精巧复杂的数学正瞄准着这个问题,说不定什么时候就会出现真[26]正的突破.

阅读文献

D.Burton 的 Elementary Number Theory(Allyn and Bacon,1980)是本出色的数论入门书(但不包括计算方面的内容).

H.Beker 和 F.Piper 的 Cipher Systems(Northwood,1982)讨论了密码学的大部分内容.虽然难读但仍值得一看. [27]

(袁向东 译)

第2章 集合,无限和不可判定性

新的地平线

有时,一个长期存在的问题的解决标志着一个数学时代或领域的结束(或开始结束),这是长期努力的顶点.更多的情况,它可以开辟一个以前可能想象不到的全新的研究领域.1963年,斯坦福大学数学家科恩(Paul Cohen)对康托连续统问题的解答正是如此.不仅这一解答的性质本身是一种新的革命,而且,为解决这一问题,由科恩发展的方法也是很新的.不久,发现这些方法可能应用于一个很广泛的领域,并且随后的20年间,看到了大量的基于科恩这一突破的活动.他的工作得到了承认,1966年,科恩被授予相当于其他科学界诺贝尔奖的数学家最高奖——菲尔兹奖章.

直到1963年以前,数学家一直面临着有两种可能性的一些数学命题正误的确定,即证明它是正确的,或证明它是错误的.实践和直觉常常成为唯一的导向,告诉人们哪些问题值得花费更大的努力,且一个不好的选择将导致人们在不可解的问题上浪费大量的时间.但是,有些人总觉得问题最终总会得到一个答案的.科恩的工作永远推翻了这种安全感.他证明了存在这样的数学命题,它既不是正确、也不是错误的,即是不可判定的.(事实上,这样说不很确切,不可判定性命题的存在在1930年已由哥德尔(Kurt Gödel)建立,但是正如我们将在后面解释的那样,对于通常的一般的数学,它并没有产生像科恩的结果那样的影响.) [28]

为了说明 1963 年发生的这场革命，有必要回到数学的本质，回到来自本世纪之交的奠基性概念上.

公 理 化 方 法

面临一些命题的正误的判定，物理学家、化学家或生物学家——事实上几乎每个科学家，通常将进行一些实验，或至少利用依靠实验证据的一些推理．事实就是如此，这些科学家中的每一位都在研究物理世界的某些方面．因而，物理世界是“一切是非”的最后仲裁者．但在数学上将会怎样呢？

在最基本的水准上，数学很像某些其他的物理科学，其中，我们周围世界的某些方面被选出来加以详细研究．所以，宇宙自身能够提供一些信息．如果你想检验一个三角形的内角和是 180°，你可以去测量许许多多三角形的角．尽管这将清楚地提供为物理学家或化学家所能接受的那种证明，然而这并不是数学家实际所做的事情，这一过程也不能构成任何三角形内角之和是 180°这一断言的一个数学证明．

“粗糙”的实验方法对于确定数学的真理是不合适的，其原因在于数学的本质，即什么是数学．尽管数学根植于物理世界，但是，数学是一门精确的、理想化的学科．点、线、面和其他理想的数学结构在现实世界并没有精确的对应物．(关于这一点，第 4 章将作一些有趣的说明．)数学所做的是从完全抽象的、理想化的观点来观察世界，和以完全精确严密的方式进行抽象推理．简单(但重要)的例子将有助于说明问题．

最基本的数学抽象之一是数的抽象，我们大多数人首先(通常很早)遇到的数学抽象就是数．当我们像孩子一样开始考虑数的时候，[29] 由于某种几乎是不可思议的过程，我们都会认识到，在三个苹果、三位叔伯、三朵花等之间有某种共同的东西．“三”的抽象导致了数 3 这一思维概念的形成．当然，这一过程不可思议之处正是在于世界上并

没有任何东西是数字3.它是一个纯粹抽象的概念,当我们说到数3或任何其他的数时,我们并没有不舒适(或甚至尴尬)的感觉,这只是因为我们已经熟悉了它.(当你试图不用词three或threeness来解释数3是什么时,你就会认识到这个概念是多么抽象.这当然是做不到的,然而我们却没人为此担心.)所有的其他数学抽象同样是如此:尽管它们也许来源于现实世界,但是,抽象自身纯粹是概念性的,离开我们的思维就不能存在.

由数的概念,我们就有了前一章所讨论的数论学科的基础.但是,你究竟怎样来进行抽象呢?不用说是要用一种被证明对物理世界是有用的方式来进行.唯一的方法就是从制定一些基本的法则开始.对于数来说就是提出对于所有数都假设成立的公设(或公理),然后,从这些最初的公设出发使用严密的逻辑的推理进行演绎.(写下这些决定逻辑推理过程的公设也是可能的,这是作为数学分支的数理逻辑的工作,与本章的主题紧密相关.)例如,我们由经验知道,当两个数相加时,它们的顺序是不重要的.如,5+3和3+5是相同的.所以,包括在算术运算中的一条合理的公理就是命题:

$$\text{对于任意数 } m,n,\ m+n=n+m.$$

这个特殊的命题就是所谓的*加法交换律*.另一个例子是*加法结合律*:

$$\text{对于任意数 } m,n,k,\ (m+n)+k=m+(n+k).$$

同样,它也来自实践中对加法运算的观察.例如,要把三个数3,5和10相加,无论是先将3加5(得8)然后将所得结果加10,还是5加10(得15)然后将所得结果加3,都无关紧要.两种情况都得到同样的结果18.

[30]

在采纳上面两个公理的过程中,我们已经迈出了一大步,这一步不是别的,而是基于信念的行为.尽管可以通过检验大量的例子来(实验地)验证这两条公理,但是,由于有无限多个数,检验每个例子理论上是不可能的.因此,当我们涉及非常大的数,比如说有上百万个数字的数时,还能相信这两条公理依然正确吗?这似乎是合情合理的,甚至可能是"显然"的.但是数学(和大多数其他学科)充满了

"明显真理"的例子，而这些"明显真理"最终却被证明是错误的.（例如，根据一般的经验，太阳绕地球运行似乎是"显然"的事实.）有效的证据只能提示这两条公理是正确的，但要明确证明这一点却不可能.它们的真实性只能作为假设.这就是这种假设被作为公理的原因，"公理"来自拉丁词"axioma"，意思是原理.（在一定的意义下，上面的两条公理能被证明，并且能够列出适用于自然数的更基础的公设，从这些公设能够演绎出更熟悉的算术规则.但这只是从信念行为倒退一步，而没有能取消它.）

为了突出上述要点，也许值得一提的是：尽管所考虑的两条规则已被作为整数运算的公理接受下来，但是，对于无限数的某些系统，交换律却是不正确的（参见下文），且结合律不适用于计算机算术.（当很小的数加上很大的数时，这个规则不成立.）

目前，数学理论的公理化发展也许值得进一步详细考查.因为我们对其某些方面已作过一些讨论，整数（所有的正、负数）运算将提供一个极好的例子.

整数：一个例子

下面的公理对于整数基本运算（即加法和乘法）的研究是合适的.

(1) 对所有 m,n，$m+n=n+m$，且 $mn=nm$.（加法和乘法的交换律.）

[31]

(2) 对所有 m,n,k，$(m+n)+k=m+(n+k)$ 且 $(mn)k=m(nk)$.（加法和乘法的结合律.）

(3) 对所有 m,n,k，$m(n+k)=(mn)+(mk)$.（分配律.）

(4) 存在数 0，使对于任意数 n 必有 $n+0=n$.（加法单位元的存在.）

(5) 存在数 1，使对于任意数 n 必有 $n\times 1=n$.（乘法单位元的存在.）

(6) 对于任意数 n,存在另外一个数 k,使得 $n+k=0$.(加法逆元存在.)

(7) 对任意 m,n,k,如果 $k\neq 0$,且 $km=kn$,那么 $m=n$.(消去律.)

从这些公理出发可以证明所有通常的整数运算的性质.例如,有一个适用于加法的类似于公理 7 的规则:

$$如果\ k+m=k+n,那么\ m=n.$$

证明这个规则,从假设 $k+m=k+n$ 出发,然后,由公理 1 得

$$m+k=n+k.$$

由公理 6,设有数 l,使得 $k+l=0$,那么把 l 加到上面方程的两边,我们得到

$$(m+k)+l=(n+k)+l.$$

这样,由公理 2,

$$m+(k+l)=n+(k+l).$$

换句话,考虑到 l 的选取,

$$m+0=n+0.$$

利用公理 4,由最后的方程立即得到

$$m=n.$$

[32] 这正是要求的.

其次,证明结果 $x\times 0=0$,对于任意的 x 都成立.我们证明如下:

$x+0=x$(由公理 4,代 $n=x$),

$=x\times 1$(由公理 5,代 $n=x$),

$=x\times(1+0)$(由公理 4,代 $n=1$),

$=(x\times 1)+(x\times 0)$(由公理 3,代 $m=x,n=1,k=0$),

$=x+(x\times 0)$(由公理 5,代 $n=x$).

所以,由刚证明的公理 7 的加法类似物(代 $k=x,m=0,n=x\times 0$),得:

$$0=x\times 0.$$

这时,你可尝试亲自证明下面的关于整数运算的一般事实.在每

个情形你都应明确只能使用已知的事实，因为它们是公理或者说已经被证明．

（1） 存在唯一的元素 0，满足公理 4 的要求：即，如果 $\varnothing$ 对于任意数 n 满足性质 $n+\varnothing=n$，那么 $\varnothing=0$．（0 的唯一性．）

（2） 存在唯一的元素 1，满足公理 5 的要求．（1 的唯一性．）

（3） 对于每一对 m,n，存在唯一的一个数 k，使得 $n+k=m$．

注意到上述最后的这个结果保证了在整数中减法总是可能的（这唯一的数 k 就是 $m-n$），即使在公理中没有提到这个运算也将如此．这个结果的一个特殊情况就是当 $m=0$ 时，证明了公理 6 中 k 的唯一性．

当然，如果数学的每件事都必须像上面那样详细证明的话，那么，数学家的工作将几乎是不可能的．数学研究之所以能够进行，是因为数学知识是积累式发展的：一旦某些东西被建立，就可以立即被使用而毋需再加证明．（这个现象的一个很简单的例子是上述 $x\times0=0$ 的证明．）因此，只是在以公理为基础的理论发展的最初阶段，这种详细的证明才是必要的．因此，大部分的数学推理更像是其他科学中使用的“通常逻辑”的严格翻版． [33]

相容性、完备性和真实性

当代数学的内容由公理推演出的结论组成．这些公理和物理世界的任何东西都没有关系（直接的或间接的）．前节给出的关于整数运算的公理是由我们所熟悉的整数的加法和乘法运算的行为检验得到的．（这里我们所熟悉的整数只意味着整数的很小部分，应该记住仅仅是在 18 世纪负数才被广泛接受，参见第 3 章．）但是，一旦公理被确定，所有关于它们的“普遍真理性”，即这些公理涉及何种具体对象的问题，都变得无关紧要．例如，前节给出的公理中，任何地方都未提到一个数到底是什么．事实上，有很多其他的数学对象也已被证明（作为形式定义的结果）满足这些公理．因为公理系统常应用于很多

不同的情况.数学家们往往引入一些名词来描述满足一个特殊公理系统的结构.任何满足前述公理系统的数学结构都称为一个整数域(如果公理7被取消,则称为环).因此,为了说明整数(连同它们的加法和乘法的算术运算)满足这些公理,只须说它们构成一个整数域就够了.有理数(分数)、实数和复数提供了整数域的另外例子.

从给定的公理系统出发由逻辑推理证明的任何结果对于满足这个公理系统的抽象结构都将是正确的.但是,关于这些结果在现实世界中的真实性问题不仅不能回答,而且根本没有意义.如果这些公理很好地反应现实世界中的某些现象,那么,这些公理的推论也将很好[34]地反应现实世界,甚至可能提供一些对人类有益(或可能导致毁灭)的有用信息.但是,就数学本身而言,最初的公理是否与现实世界有关联并不重要.一些已经导致很有趣的数学的公理系统似乎和物理世界没有什么关系,但这并不是说将来不可能发现有某种关系!数学家使自己(程度不同地)与现实隔离起来,以此为代价,他们就能够在一个绝对确定的世界中进行研究,同时具有使研究结果获得广泛应用(首先是在数学内部)的潜力.因为他们的公理系统不仅适合于他们头脑中(可能)存在的结构,而且适合于其他的结构.

如果"真实性"不是尺度,那么,究竟是什么样的考虑支配着一个公理系统的形成呢?

一个重要的要求是相容性:不能从公理系统推出两个相互矛盾的结果.所有的公理系统都必须满足这个要求,尽管相容性的证明不仅非常困难,而且还有哲学问题的困扰(正如后面将要说明的那样).

另一个要求就是完备性,它关系到任何试图刻画某种特殊数学结构(如整数运算)的公理系统.这个公理系统应当充分丰富,以致于能够证明关于该结构的所有"真的事实".同时满足上面两个要求是一种微妙的平衡.要达到完备性,可能需要引进越来越多的公理.然而,公理越多,导致不相容的可能性也就越大.

哥德尔不完备性定理

在本世纪之交，世界著名的德国数学家希尔伯特(David Hilbert)在公理化方法的严格形式化范围内提出了一个数学发展的计划.根据希尔伯特的信念，所有的数学都能被看作是以指定公理为基础的符号的形式的逻辑操作.(原则上，这将意味着可以用编有程序的计算机去研究所有的数学.)但是，1930年，年轻的奥地利数学家哥德尔通过两个完全出人意料的定理证明了希尔伯特的计划是不可能实现的.

哥德尔证明了：任何足以包含初等算术的相容的公理系统，一定 [35] 存在着与此公理系统相关的命题，从这些公理出发，该命题既不能被证明，也不能被否证(第一不完备性定理).并且，这些不能证明(从公理出发)的命题与公理系统是相容的.这样(对于希尔伯特计划来说)，十分重要的相容性概念就注定成为永远不可捉摸了(第二不完备性定理).

尽管哥德尔的结果意味着公理方法不能雄踞像希尔伯特想象的那样包罗一切的地位，但也不应认为它宣告了公理方法在它过去和现在一直被应用的通常数学中的死亡.相反，本世纪已经看到公理方法在数学中所占据的绝对优势.哥德尔迫使我们放弃的是这样的信念或希望，即一个公理系统能够解答我们可能向它提出的所有合理的问题.

事实上，随着公理方法的越来越成功，在哥德尔宣布其结果以后的年代，有一种信念在日益增长，即只有那些非常特殊的命题才是不可证明的.例如，哥德尔的第一不完备性定理是通过证明初等数论中可能提出这样的命题而获得的，这种命题类似于(英语中)明显的自相矛盾的断言：

本页方框内的句子是错误的

(在哥德尔的数论类似物中，"不可证明的"代替了"错误的".由于这类命题的陈述需要用到初等算术，因此哥德尔的结果就只能被应用

于可提供这种算术的公理系统.但任何要实现希尔伯特计划目标的公理系统当然都必须使你能进行初等算术运算!)

回到哥德尔的第二不完备性定理,尽管对于一个公理系统相容性的考虑十分重要,但是,它能不能用这些公理来证明却并不那样重要.一开始写下公理的时候,数学家们就有一种默契,即假定这些公理是相容的,而他们的主要兴趣则在于使由公理推出的结果具有更可靠的性质.例如,在数论中,公理的相容性并不是数论学家们普遍关心的问题.也许,哥德尔的不完备性结果并不是真有那么重要,至少在一个自信的年轻美国人成功地证明相反的结论之前,人们认为情况似乎是这样的.科恩1963年获得的引人注目的结果摧毁了不完备性不影响“实际”问题的舒适感觉,并且事情是发生在数学的最基本、最重要的部分——集合论中.

[36]

公理化集合论

在这个世纪之交,19世纪抽象的纯粹数学的发展导致了德国数学家康托(Georg Ferdinand Ludwig Philipp Cantor)的工作,他系统地描绘了一个能够为全部数学提供基础的通用的数学框架.康托创立的这个学科一直为我们提供着良好的基础.这个学科就叫做集合论.它的概念和方法已有效地渗透到所有的现代数学.但是,自从康托最初的陈述以来,集合论已经历了日新月异的迅猛发展.下面将介绍这方面的发展.然而,在此之前介绍一些形式逻辑的知识是必要的.在康托发展他的集合论思想,特别是用以衡量无限集“大小”的无限数系的同时,弗雷格(Cottlob Frege)正在创立现在所谓的“谓词逻辑”理论.广义地说,谓词逻辑提供了一个足以表述任何数学概念的通用的形式化语言,其重要性并不在于数学家们有应用谓词逻辑进行工作的强烈需要和愿望.事实上,由于这种语言的简单性,在大多数情况下,弗雷格框架内数学概念或论证的表述是相当冗长繁琐的.弗雷格工作的重要性首先在于它清楚地证明了很多数学分支都是一个协调

的整体的一部分，其次（更重要的）是能够对构造性证明中数学家所使用的演绎方法作出恰当的分析.（虽然应该注意目前在试图发展计算机程序以获得或帮助获得数学结果的努力中，人们看到谓词逻辑在表述数学概念和证明方面的用处正日益增长.显然，为了使数学能够具有适合于计算机处理的形式，就必须使用一种精确简单的语言，谓词逻辑恰好提供了这样一种框架.）[37]

康托使用的集合概念是相当简单的.集合是对象的任意收集，或者至少是数学对象的任意收集.关键是将这种收集本身也看作是单个的对象.小的有限集可以通过列出它们的成员（或元素）来表述，通常被放在大括号之间.

$$\{1,3,5,9\}$$

表示元素是数 1,3,5 和 9 的集合.对于较大（并可能无限）的集合，不可能列出它的所有元素，那么，必须依靠一些特征来确定它是什么集合.所有的具有性质 $P(x)$ 的对象 x 的集合的标准记法是：

$$\{x \mid P(x)\}.$$

所有的素数集合（一个无限集）可以记作

$$\{x \mid x\text{ 是素数}\}.$$

也有一些集合没有明确的定义性质，这种集合不能借助它们的元素来说明，但是，这和目前的初步的讨论并没有实质性的关系.粗略地说，这种集合是"缺席"产生的，因为收集的概念并不需要确定集合的性质的存在——但是，这方面的讨论相当高深而微妙.我们现在不谈这些难以理解的集合，主要讨论可以由性质确定的集合.那么，在集合的构造中，什么样的性质是允许的呢？可能像你现在所期望的那样，最初的回答是任何可以由弗雷格的谓词逻辑表述的性质.由于其语言的形式化性质，这一定义是精确的，同时又适用于数学中遇到的所有性质.

到现在为止，事情似乎仍不容乐观.集合论提供了一个适当的框架，所有的数学对象和结构都可能以它为基础而构造出来，并且，弗雷格的谓词逻辑提供了一种可以定义、讨论这些对象和结构的一般

语言,包括基本的集合概念本身.弗雷格在他的两卷著作《算术基础》[38]中广泛地使用了集合论的概念,打算作为他一生工作的顶峰.

正当弗雷格著作的第二卷付印之际,1902年6月16日,他收到了著名的英国逻辑学家罗素(Bertrand Russell)的一封信.信中开始赞扬了弗雷格的第一卷,接着,罗素,就转到了他这封信的要点.他说:“只是在一个地方我遇到了困难”,随后就是对于他一年前所做的一个精确观察的简洁说明,这个观察彻底地推翻了弗雷格的全部理论.

众所周知,罗素悖论非常简单,同时又非常深刻.根据康托集合论的基本原则,如果 $P(x)$ 是任何一种适用于数学对象 x 的性质,那么,存在一个对应于所有这样的 x 的集合,对这些 x 而言 $P(x)$ 成立.这就是集合

$$\{x \mid P(x)\}.$$

这里讨论的对象 x 本身也可以是一个集合,因为一个集合就像其他数学对象一样也是数学的对象.(事实上,当集合论被作为数学的基础时,每个数学对象结果都是这样或那样的一个集合.)现在对于性质 $P(x)$,罗素取命题(应用于集合 x)

$$R(x): x\text{ 不是 }x\text{ 的成员}.$$

(集合成员的标准符号是 $\in$,所以 $x \in y$ 就意味着 x 是 y 的一个成员,非成员关系记作 $x \notin y$,因此,罗素的性质 $R(x)$ 可以记作 $x \notin x$.)

设由性质 $R(x)$ 确定的集合为 y,则

$$y = \{x \mid x \notin x\}.$$

由于 y 是一个集合,所以询问 y 是否是它自身的一个成员是完全合理的.如果它是,那么 y 必须满足它自身定义的性质,也就是说 $y \notin y$,即 y 不是它自身的一个成员.另一方面,如果 y 不是它自身的一个成员,那么 y 不满足它自身定义的性质,所以 $y \in y$ 必须成立,即 y 是它自身的一个成员.于是我们便面临着一种明显的矛盾局面,这里,如果 y 是它自身的一个成员,那么可以推出它不是,而如果 y 不

是它自身的一个成员，则又可以推出它是一个真正的悖论.

罗素悖论之所以如此地具有破坏性，是因为它的绝对的简明. 它只用到了几乎所有数学都赖以生存的最基本的概念.

德国数学家策墨罗(Ernst Zermelo)提出了一条摆脱罗素悖论引起的困境的道路，他关于积分方程(一个应用性很强的数学领域)的著作引导他考虑了关于无限集性质的更深刻的问题. 1908 年，为了给自己的著作建立一个可靠的集合论框架，策墨罗发表了一篇论文，在这篇论文中他发展了一个集合论的公理系统. 随后，弗兰克尔(Abraham Fraenkel)对这个公理系统进行了修改，作为抽象集合理论的一种"正确"的公理化方法，策墨罗－弗兰克尔集合论逐渐地被接受下来.(对于涉及公理的适当的起因分析和内容解释需要更多的篇幅. 有一个基本的解释可以在各种教科书中找到——详见下文.) [39]

利用哥德尔不完备性定理当然还不足以证明策墨罗－弗兰克尔集合论的公理是相容的. 但是它们确实回避了像罗素悖论这样的悖论. 大多数数学家深信它们决不会引起任何矛盾. 由于这个理论已经被证明能经受时间和广泛应用的考验，上述信念变得越来越坚定了.

相容性如此，完备性又如何呢? 不完备性公理还告诉我们存在着一些与集合有关的命题，在所选定的公理的基础上，它们既不能被证明也不能被否证. 这一缺陷由于集合论的特殊的、基本的性质而显得非同一般. 由于现在数学的整个大厦被看作是(并且很大程度上也确实是)建筑在集合论基础之上，集合论内的缺陷可能导致数学其他领域内的缺陷. 尽管一直存在着这种可能性，但策墨罗－弗兰克尔公理却似乎足以提供一个对数学来说是适用的集合理论. 并且大多数工作着的数学家对这一危险都视若无睹，就好像对他们毫无影响一样. 直到 1963 年，科恩的突破性的工作才使这个问题充分曝光.

尽管科恩的发现产生出许多的结论，但是，它最初涉及的是一个包含康托无限数的问题. 一旦策墨罗－弗兰克尔公理被明确表述，无限数的理论就变得完全合理(这是众所周知的). 所以，我们现在再回到无限来，看一看康托的理论. [40]

无 限 集

尽管我们生存的世界是有限的，但是，为了研究它，所需要的数学却几乎处处都涉及无限. 所有自然数的集合是一个无限集，数 π 的精确表示需要无限多位小数，哪怕是很小的线段所包含的点数也是无限的，等等. 尽管人们一直努力避免无限的使用，但所产生的数学却是令人难以置信地繁复庞大. 尽管十分抽象，无限的世界却是一个十分简明的领域. 从有限进入无限很像在电视屏幕前由近往远倒退一样，当你退到足够远时，屏幕上大量模糊复杂的小光点看起来就变成清晰连续的画面. 进入到无限时，大的有限的复杂性就消失了. 这种现象不单出现在纯粹数学中. 例如，在经济学中，研究含有无限多商人的理想化经济就优于现实世界大有限经济的研究. 在物理学中，无限容积被用于探讨某些热和电能的精细概念.

无限数系及其算术的发展形成了康托关于集合论的先驱性工作的最高成就. 但是，你可能要问："我们为什么需要无限数呢？"回答是：和我们需要有限(整)数的原因相同——计数集合成员的个数. 自然数可以使我们度量一个有限集的大小. 为了度量一个无限集的大小，无限数是必须的(从中你可以预料仅仅简单地说一个集合"无限"是不够的). 接受了这一点，接着你也许要问："什么是无限数？"对这问题一个很好的回答就是："什么是有限数？"像我们在本章开始所看到的，自然数只是我们想象的虚构事物，因此，假设无限数的存在就没有什么两样. 重要的是这些无限数的性质，这正是理解康托无限数的关键.

自然数是由有限集抽象而来的(数学的集合或现实生活中的集合，如苹果的集合、人的集合等等). 数 3 是所有有三个元素的集合所共有的性质. 这表面上看来像是一个循环定义(因此，这当然就不能成为定义)，但是康托发现并非如此. 相反，在我们定义数之前，必须有像我们将要解释的两个集合"大小相等"的概念.

[41]

两个集合 A 和 B 有相等的大小，如果 A 的每个元素恰好对应于

B 的一个元素，且反之亦然. 例如，集合

$$A = \{1,2,3,4\}, B = \left\{100, \pi, \sqrt{2}, \frac{1}{2}\right\}$$

有相等的大小，这可由以下的配对方法来检验(还有其他的配对方法)：

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ 100 & \pi & \sqrt{2} & \frac{1}{2} \end{array}$$

类似地，集合

$$A = \{a, b, c\}, B = \{\text{脚}, \text{袜}, \text{鞋}\},$$

也可由配对证明其大小相等：

$$\begin{array}{ccc} a & b & c \\ \updownarrow & \updownarrow & \updownarrow \\ \text{脚} & \text{袜} & \text{鞋} \end{array}$$

请注意在上述两个例子中都没有提出集合元素个数的概念，了讨论“大小相等”，既不需要建立“大小”的概念，不必仅限于考虑有限集合. 同样的思想也适用于无限集(尽管在这种情况下，要像上面那样清晰地描述配对是不可能的). 然而，当应用于无限集时，你很快就将发现会遇到一些意想不到的结论. 例如，设 A 是所有自然数的集合，B 是所有偶数的集合. 直观地，B 的“大小”恰好是 A 的“一半”. 但是，根据我们的定义，这两个集合大小相同，这可以通过配对方法 [42] 来验证：

$$\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & \cdots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ 2 & 4 & 6 & 8 & 10 & \cdots \end{array}$$

这里没有矛盾——如果有的话也只是出于我们的成见. 这就是说无限集和有限集并不总是具有相同的行为方式.

希尔伯特旅馆提供了无限集行为的一个很好的例子. 这个理想

化的建筑物有无限多房间,以所有自然数1,2,3等等来编号.一天晚上,碰巧所有房间都住满了.(在这个故事中人数也是无限多.)在不撵走任何旅客的情况下,一个新来者仍然可以被安置进去.为此,只需将新来者安置在1号房间,原1号房间的旅客挪到2号房间,2号房间的旅客搬到3号房间,等等,这样让所有的旅客依次挪动一个房间,就可以让这个后来者住进已被腾空的1号房间.(事实上,安置无限多个新到的旅客也是可能的,你能试试怎样安置吗?)尽管这个无限的旅馆的想法也许有些牵强,但是整个论证的内部逻辑并没有任何错误.无论多么与直觉相背,这就是当你开始探索无限领域时必然会遇到的情形.

自然数和偶自然数的例子也许会引导你猜想所有的无限集大小都相等,这将意味着无限数系是不必要的.事实上,在数学上遇到的大量的无限集大小相等.例如,素数集、自然数集、整数集和有理数集都有相等大小.(和自然数集有相等大小的集合通常被称为是可数的,因为与自然数的配对提供了一个计数它们的元素的方法.)但是,康托发现,并非所有的无限集都有相等的大小.事实上,有一个越来越大的无限集的完整(无穷)的等级.康托对于这个关键事实的证明既简单又优美,证明仅使用了集合论的最基本的概念,但它是高度抽象的.因此将留到本章的最后来介绍,以便那些过于拘谨的读者可以[43]略而不读.在目前的情形下,我们仅指出实数集的大小和自然数集是不相等的(尽管它和平面上的点集与三维空间的点集有相等的大小).

无限数和康托连续统问题

一旦你领会了相等大小的概念,你就可以进一步发展一个被用来度量任何集合"大小"的数的系统,而无论这集合是有限集还是无限集.当然,"数"自身是抽象的,重要的是,如果两个集合大小相等(也就是如果它们的元素能够像上节所描述的那样配对),那么,它们

的大小(即每个集合中元素的数目)确应是相等的.例如,当你度量两个集合

$$\{a,b,c\},\{\text{Fred},\text{Elsie},\text{Fido}\}$$

的"大小"时,你发现两个集合有相同的元素个数,也就是3.另外,当你度量两个无限集

$$\{1,2,3,4,5,\cdots\},\{2,4,6,8,10,\cdots\}$$

的"大小"时,你又一次发现这两个集有相同的元素个数——在这情形,这个"数"是最小的无限数,记作$\aleph_0$(沿用康托的符号).(读作"阿列夫-零"(aleph-null),aleph是希伯来语alphabet的第一个字母.下标为0的原因是很清楚的.)

那么数3代表什么?它是所有三元集的共性.或者,换一种方式说,它是所有和集合$\{a,b,c\}$大小相等的集的共性.这里3是一个来自大小相等概念的抽象物.可以用不同的数学方式来精确地叙述这一命题,这里略而不谈.如果你愉快地(果真如此?)接受了"数3"的概念,那就应该能同样愉快地接受数"$\aleph_0$".它是和所有正整数集合有相同大小的一切集合的共性.

44

如前所说,并不是所有的无限集都有相同的大小——有一个完整(无穷)的无限数等级.就像有一个有限数的无穷序列1,2,3,…一样,也存在一个无限数的无穷序列$\aleph_0,\aleph_1,\aleph_2,\aleph_3,\cdots$,其中每一个都比前一个"大".

康托的$\aleph$数的加法和乘法异常简单(这乍看有点令人惊奇),在每种情况结果都是两个无限数的较大者.例如

$$\aleph_0+\aleph_1=\aleph_1,$$

$$\aleph_1\times\aleph_3=\aleph_3.$$

(希尔伯特旅馆对应于事实$\aleph_0+1=\aleph_0$,要使旅馆爆满,必须有$\aleph_1$个旅客到来.)

数学中出现的许多无限集的大小为$\aleph_0$.例如,所有正整数的集合、所有整数(即正的和负的)的集合、所有有理数集合和所有素数的集合,大小都为$\aleph_0$.但是,正如康托证明的那样,所有实数集合的元

素肯定就比$\aleph_0$多.这立即引起问题:这个集合的大小是什么呢?因为它不是$\aleph_0$,所以必定是$\aleph_1,\aleph_2,\aleph_3,\cdots$中的一个,但是是哪一个呢?尽管进行了很多尝试,康托尔还是不能回答这个似乎很简单的问题.很多其他一流数学家也都相继失败了.事实上,这个以康托连续统问题而著称的问题的许多求解努力最后都以失败而告终,以致希尔伯特在1900年巴黎国际数学家大会上所作的主要演讲中,将这一问题列为在新世纪到来之际数学家面临的一系列最重要的挑战之一.(因为要求实连续统的大小,这问题的名字因此而起,实连续统这个词被用来描述实数的集合,这时实数被看作组成实直线的点.)

1938年,这一问题取得了一些进展,哥德尔使用数理逻辑的新方法证明了从策墨罗-弗兰克尔公理出发,不可能证明实数集合的大小不是$\aleph_1$.但是,这并没有解决连续统问题,因为这些公理对于用这样或那样的方式来解决问题很可能并不充分.

尽管有这种可能性,然而,在哥德尔得到其结果的随后几年,有一个普遍的希望,就是在策墨罗-弗兰克尔框架之内连续统问题仍然是可以判定的.在这种情况下,由于哥德尔已经指出了我们不可能[45]证明实数集合的大小不是$\aleph_1$,所以,连续统的大小必需(假设)是$\aleph_1$.并且随着时间的推移,它将最终获得确切证明.因此,当数学的某一部分需要关于连续统大小的知识时,假设这个预期的结果似乎并不是完全不合理的,并且,大量的数学结果都是在假定"连续统假设"(仅仅作为假设)成立的情况下被证明的.

然后,在1963年,斯坦福大学的科恩发展了一种新的逻辑方法,利用这种方法,他证明了连续统假设不能由策墨罗-弗兰克尔公理推演出来.结合较早的哥德尔的结果,这就证明了在策墨尔-弗兰克尔系统内,连续统假设事实上是不可判定的.

接下去怎么办?对于由科恩的结论带来的事态有两种看法.一种结论是它证明了策墨罗和弗兰克尔的公理是不合适的.并且根据这种看法,问题显然十分严重.知道这一公理系统像哥德尔不完备性定理所预示的那样存在缺陷是一回事,但这个系统不能解决像"有多

少实数?"这样一个基本问题那又是另一回事了,它是对于这个理论的致命的打击.一些数学家对这一打击的反应是建议为克服这个新发现的缺陷而提出一些附加的公理.但是,如果采取这种路子,你将面临必须寻找适当的附加公理的任务.由于集合论的简单的本质和它在数学中的基础地位,你引进的任何公理必须是"可相信的",这意味着,即使我们不是一看就明白(有些策墨罗－弗兰克尔公理为了明白它们的意思就颇费思考),至少当开始研究这些公理时,它们的出现必须是很自然的.(这种考虑使人们不可能采取简单地将连续统假设本身作为一条集合论公理的捷径——有什么正当的理由可以这样做呢?)半个多世纪以来,数学家们一直使用公理化集合论进行工作,没有遇到过任何这种附加"原理",这一事实导致了大多数的专家们认为事实上并没有"丢失公理".

剩下的选择就是从科恩的发现得出另外的结论,即:无论你多么不愿接受,但集合论不只有一种,而是有若干种.(就像 19 世纪人们认识到不只有一种"正确的"几何学,而是有三种可选择的几何学,每一种都有它自己的明确的性质和结果.)在一些集合论中,连续统假设是正确的,在另外一些中,它将是不成立的. [46]

请注意上面我们使用了"几种集合论"这样的措辞,而不是说只有"两种".因为由策墨罗－弗兰克尔公理不可判定的不仅仅是连续统假设.在科恩 1963 年的最初发现之后,很明显,他的新方法(力迫法)不只是可以用于集合论自身,而且可用于很多其他的情况.随后的 20 年间看到了数学中许多经典的未解决问题不可判定性的证明.以前那种认为只要有足够的时间和精力,任何"真正的"数学问题都能用这样或那样的方法获得解决的希望彻底破灭了.除了正确的命题和错误的命题之外,还有第三种不可判定类型的命题,这种命题既不是正确的也不是错误的.至少科恩的力迫法提供了证明一个问题属于第三类型的一种方法,所以,他的结果确实对数学起到了积极的贡献.

康托的证明

康托如何证明存在一个无穷的完整等级呢？对于愿意看到一个高度抽象的数学推理范例的读者来说，这里给出康托论证的一个现代的描述.它以集合论的子集概念为开端.

如果 X 是任一集合，任何由取自 X 的元素组成的集合，都称为 X 的子集.这样，集合

$$\{a,e,d\}$$

是集合

$$\{a,b,c,d,e,f\}$$

的一个子集，素数的集合是所有整数集的一个子集.

现在考虑由集合 X 的所有子集组成的集合，这样一个集合存在吗？回想一下罗素悖论就足以提醒你必须谨慎对待集合存在的假定.在目前情况下这是没有问题的(如所周知).策墨罗－弗兰克尔集合论的公理之一保证了存在这样一个集合，它称为 X 的幂集.记为 $P(X)$.例如，如果 $X=\{a,b\}$，那么 $P(X)$ 由集合

[47]

$$\varnothing,\{a\},\{b\},\{a,b\}$$

组成.符号 $\varnothing$ 是什么呢？它表示空集(或零集)，没有元素的集合.如果它确实是一个集合，那么它将是任何其他集合的一个子集.因为按照一种平凡但仍然有效的逻辑，一个没有元素的集合具有这样的性质：它的所有元素都属于你选择的任何集合 X.仅仅基于这样的推理，你也许仍会认为把空集概念与其他数学"虚构物"放在一起并不是一个好主意.但若按同样的理由，那么零也不能被看成是一个合理的数.这样，你对于为什么空集也和数学中所有其他集合一样被看成是真正的集合就会有一些认识了.它是一个像数0那样的"零元素".(事实上，0是集合 $\varnothing$ 中的元素数目.)

另一个例子：如果 $X=\{1,2,3\}$，那么 $P(X)$ 的元素是集合：

$\varnothing,\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}$.

上面的两个例子应足以说明 $P(X)$ 似乎是比 X 更大的集合. 当 X 有两个成员时, $P(X)$ 有四个成员, 当 X 有 3 个成员时, $P(X)$ 有 8 个. 事实上, 容易证明, 如果一个有限集 X 有 n 个成员, 那么 $P(X)$ 有 2^n 个元素. 从第 1 章关于指数函数 2^n 的讨论, 你将明白当有更多的元素加到 X 中时, $P(X)$ 的大小是如何迅速增长的. 这种增长率的差别可以延伸到无限(然而, 正如希尔伯特旅馆的例子所表明的, 这种从有限到无限的延伸并不是不言而喻的). 康托正是以此来证明有无穷多的无限的. 他证明, 对于任意集合 X, $P(X)$ 一定和 X 有不同的大小(因此是比 X 具有更大的数量级). 由此容易得到无限的无穷等级. 假设 X 是自然数集合, 那么 $P(X)$ 是比 X 更大的无限数的集合. 又 $X_2 = P(X_1)$ 比 X_1 更大, 类似地 $X_3 = P(X_2)$ 比 X_2 更大, 等等.

为了证明康托的结果, 假设 $P(X)$ 和 X 有相同的大小, 我们的目的是从这个假设推出矛盾. 如果所使用的论证逻辑上是可靠的, 那么, 必然的结论就是最初的假设是错误的(因为一个正确的假设不能导致一个错误或矛盾的结论). 这里就采用这种方法. 由于假设 $P(X)$ 与 X 有相同的大小, 那么在这两个集合之间存在一个对应. 即对于 X 的任一元素 x, 在 $P(X)$ 中存在一个相对应的元素——x 的配偶, 且它不与 X 中的任何其他元素对应; 同时 $P(X)$ 中的每个元素都是 X 中某个元素的配偶. 因为这个论证应当对于任何有限的或无限的集合 X 都有效, 因此, 就不可能像我们前面做的那样, 用箭头来表示这种对应——但请参阅框图 A. [48]

现在考虑 X 的任一元素 x. 那么, 它的配偶 A 是 $P(X)$ 的一个元素, 即 A 是 X 的子集. 所以 A 是由 X 的一些元素组成. x 自身在这些元素中吗? 也就是 x 是 A 的元素吗? 这是一个完全合理的问题. 对于 X 的某一元素 x, 回答可能是"是"或"不是". 设 U 是 X 中所有不是它自身的配偶的元素 x 的集合(这能使你想起些什么吗? 见下文). 集合 U 由 X 的元素组成, 因此 U 是 X 的一个子集, 也就是说 U 是 $P(X)$ 的一个元素. 因此, U 应是 X 的某个元素的配偶, 记此元素为 w($U = w$ 的配偶).

现在问:"w 是 U 的元素吗?"如果是,那么 w 必须满足 U 的定义,就是说 w 不是其配偶(在此情形中即 U)的元素;另一方面,如果 w 不是 U 的元素,那么,w 不满足 U 的定义,所以 w 是其配偶 U 的元素.这是站不住脚的——我们得到了一个矛盾.如前所述,唯一可能的结论就是最初的假设即 $P(X)$ 和 X 有相等大小必定是错误的.因此 X 和 $P(X)$ 大小不等.康托的结论得证.

这是一个化祸为福的精彩例证.你可能已经注意到上面的讨论和罗素悖论之间的相似性.但是,在这种情况下,所有的步骤都可以从策墨罗-弗兰克尔公理中找到根据.你得到的不是毁灭整个理论的悖论式的矛盾结果,而是正中下怀的最初假设的否定.

集合论的发展是现代数学的巨大成就之一,当今,几乎没有哪一个数学分支不或多或少地受到集合论思想和方法的影响.罗素曾盛赞康托开创这一学科的功绩,认为"可能是这个时代最值得夸耀的伟大成就",希尔伯特也说道:"没有人能从康托为我们建造的乐园中把[49]我们赶出去."对这些观点是不会有人持异议的.

框图 A:康托定理的证明

给定一个无限集 X,求证 $P(X)$的大小与 X 不同(因此更大).证明的思路是假设 X 的元素和 $P(X)$的元素之间存在一个配对,然后导出矛盾.如果用字母来标记 X 的元素(当然,这些字母构成一个有限的但对目前的解释却足够使用的集合),那么,这个假设的配对的一部分可能如下:

X 的元素　　　X 的子集

$$y \longleftrightarrow \{a,b,c,d\} = A(y)$$
$$z \longleftrightarrow \{b,d,p,q,z\} = A(z)$$

在此情形,y 不是其配对集合$A(y)$的元素,而 z 是其配对集合$A(z)$的元素.设 U 是X 中所有那些不是其配对集元素的元素所组成的集合:

$$U = \{x \mid x \notin A(x)\}.$$

由于 U 是 X 的子集,所以必有某个元素 x 与之配对,设为 w:

$$w \longleftrightarrow U = A(w).$$

对于 w 是否是集合$A(w)$的元素的考查将导致矛盾.如果 w 是$A(w)$的元素,那么,这将意味着 w 不在 U 中,但由于 U 就是$A(w)$,这就得出矛盾.另一方面,如果 w 不是$A(w)$的元素,那么 w 将是 U 的元素,同样因 $U = A(w)$,也得出了矛盾.[50]

阅读文献

介绍数理逻辑的书有 J.N.Crossley 的 What is Mathematical Logic?（Oxford University Press，1972）和 C. W. Kilmister 的 Language，Logic and Mathematics（English Universities Press，1967）．还有一系列其他的书．Wilfred Hodges 的书 Logic（Pelican，1977）主要涉及这个学科的非数学方面，但也值得一读．

对于包含本章所提到的各个主题的集合论的一个优美完整的介绍是 Keith Devlin 写的 Fundamentals of Contemporary Set Theory（Springer－Verlag，1980）． [51]

（包芳勋译）

第3章 数系和类数问题

有180年历史的问题的解决

1983年,马里兰大学和波恩马克斯·普朗克研究所的唐·蔡格尔(Don Zagier)与布劳恩大学的(在罗德岛的教会学院)格罗斯(Benedict Cross)宣布,他们解决了高斯(K.F.Gauss)于1801年提出的一个著名数学问题——类数问题.虽然他们的证明在数学中还不是最长的(第5章将讨论那个最长的证明),但它已超出了大多数数学证明,足足有300页.不过数学家们感到新奇的并不是证明的长度,而是这个非常间接的证明用一种不同寻常的方式把两个表面上毫无关系的数学领域联系了起来.

这个问题和它的解决虽然都是高度抽象的,而且涉及某些很难的数学内容,但本质上都与某种数系有关,因此我们就有可能讲清其中一般性的问题,这就是本章的目的.同时我们还要追溯今日数学的[52]许多历史背景.让我们首先来看看:

163不同寻常的特性

18世纪,伟大的瑞士数学家欧拉(Leonhard Euler)发现了(不知是怎么发现的)公式

$$f(n)=n^2+n+41$$

有一个相当奇怪的性质:让 n 等于一个0到39之间的数,$f(n)$的值

总是素数．比如，$f(0)=41$ 是素数，$f(1)=43$，$f(2)=47$ 也是素数．尚未发现其他的二次式可以产生这么多素数（从 $n=0$ 开始，依次取 n 的值所产生的素数）．虽然这样产生的素数序列到 $n=40$ 时中止了，因为 $f(40)=41^2$，但此式还是能产生大量素数．在前一千万个值中有近三分之一是素数，这比任何二次式产生的素数都多（第 6 章将进一步讨论素数生成公式）．

欧拉的公式在制造素数方面显得与众不同，里面大概有些特殊的奥妙吧？是的，整数公式的性质往往跟形式相同的实数（甚至复数）公式的性质非常接近．实际上，有一门叫做解析数论的数学分支学科就是利用这种现象（见第9章）进行研究的．那么欧拉公式当被看成是实数公式时有些什么性质呢？

首先，让我们用 x（通常用于表示实数的符号）代替 n（通常用于表示整数的符号）重写上述公式：

$$f(x)=x^2+x+41.$$

记得中学代数的人都会想到二次方程

$$ax^2+bx+c=0,$$

已知 a,b,c 的值时，可以解出 x 的值，甚至还能记起一个公式，它给 [53] 出方程的两个解：

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

（加或减号的两种可能给出方程的两个解）．

因为负数不能开平方（在实数范围内），b^2-4ac 的符号就是很重要的，如果为正，则方程有两个根；如果为负，则没有实根（特殊情况，如果为 0，只有一个根）．b^2-4ac 叫做二次方程的判别式．

欧拉二次式 x^2+x+41 的判别式是什么呢？它的 $a=1,b=1,c=41$，故

$$b^2-4ac=1-164=-163.$$

因判别式为负，我们马上得知二次方程

$$x^2+x+41=0$$

没有实数解(两个复数解是:

$$x=-\frac{1}{2}\pm\frac{1}{2}\sqrt{163}i.$$

请看本章后面对复数的讨论).

欧拉公式作为一个"素数生成器"的特殊表现的背后蕴含着某种原因,信不信由你.这并不是由于判别式是负数(许多公式的判别式都是负数),而是由于它的值是 -163.你也许会问:"163又有什么特别的?"读下去你就会发现这的确是一个很特别的数,它与一些基本的数学常数紧密相关.

数学中最常见的特殊"常数"——它们往往在最料想不到的地方抛头露面——是些什么数呢?最明显的一个是 π——圆的周长与直径的比.(这一定义已表明 π 是特殊的.为什么不管圆的大小,对每个圆来说这个比值都是一样的呢?)

[54]

19世纪后期,人们知道了 π 是无理数,就是说它用十进小数表示时不会终止或出现某种循环模式.取 π 的20位小数值是:

$$\pi=3.14159265358979323846.$$

现在计算机已经把 π 的值计算到了3千万位以后.

除了依据圆来定义外,π 还出现在其他许多地方.比如,无穷数列

$$1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\cdots$$

的和的值是 $\frac{\pi}{4}$(19世纪数学的杰作之一是研究这种无穷和的方法的进展).还有,

$$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}+\frac{1}{25}+\cdots$$

(此数列中的第 n 项是 n^2 的倒数)的和为 $\frac{\pi^2}{6}$.

π 还意外地出现在如下情形中:在一个布满平行线的木板上投掷一根火柴,其中相邻平行线间的距离为一根火柴杆长,则火柴落在木板上后碰到某一条平行线的概率正好是 $\frac{2}{\pi}$.

下一个最常见的数学常数是 e，即自然对数的底．像 π 一样，e 也是无理数，它的十进小数表示是无穷无尽的．取 e 的20位小数值是：

$$e = 2.71828182845904523536.$$

同 π 一样，e 也有多种定义方式．比如：e 可看成这样的数，以它构造的函数

$$y = e^x$$

的图像在任何一点的斜率等于在那个点处的（y 的）值．又例如，若人口 p 依照规律

[55]

$$p = e^t$$

增长（t 为时间），则在任一时刻的增长率正好等于此刻的人口数．

e 的另一个相关的定义为：它是这样的数，使得由曲线 $y = \frac{1}{x}$，$y = 0$，$x = 1$ 和 $x = e$ 所围成的面积恰等于1（见图3）．用积分式表达，e

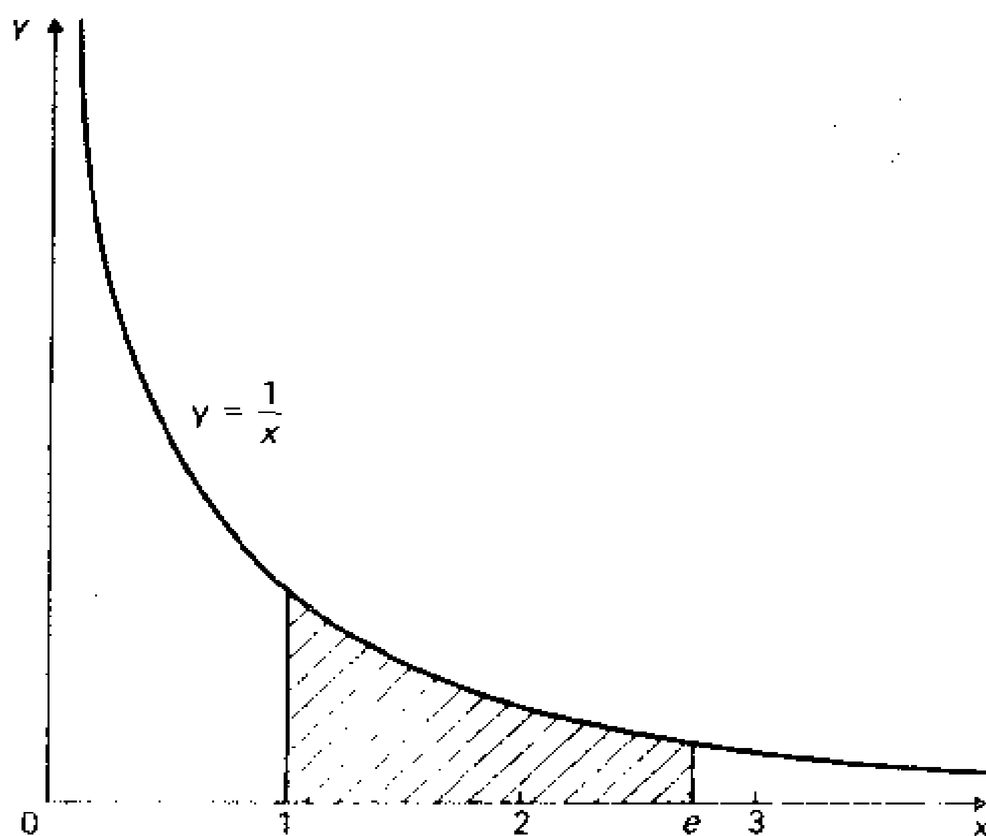


图 3　数学常数 e 作为使阴影部分面积正好为1的数的定义．

的这一定义就是：

$$\int_1^e \frac{1}{x}dx = 1.$$

还有一个定义涉及无穷和

[56]

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots,$$

其中 $N!$（读作"N 的阶乘"）代表 $1\times2\times3\times\cdots\times N$ 的积. 实际上这是公式

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

的特例.

这里先提一下在本章后面要讨论的有关复数的话题，即使 x 是复数，上面的公式也是成立的（就是说，x 为形如 $a+bi$ 的数，其中 $i=\sqrt{-1}$）. 这将引出一些令人吃惊的结果. 比如，欧拉发现了

$$e^{\pi i} = -1.$$

换句话说，当无理数 e 的指数为虚数 $\sqrt{-1}$ 的 π 倍时，其值是整数 -1. 另一个涉及 e，π 和 $\sqrt{-1}$ 的令人吃惊的结果是：

$$i^i = e^{-\frac{\pi}{2}} = 0.2078795763\cdots.$$

现在我们要进入数学常数这一讨论的要害之处了. π，e 和 $\sqrt{163}$ 三个数都是无理数，而

$$e^{\pi\sqrt{163}} = 262537412640768744.000000000000$$

（取 12 位小数）. 事实上，此数并非整数，更精确的值是

$$262537412640768743.999999999999250$$

（精确至 15 位小数）. 这个值非常接近于整数，而对其他自然数 k，没有比 k 为 163 时 $e^{\pi\sqrt{k}}$ 的值更接近于整数的. 令人吃惊的是这一特殊情况又发生在 163 这个数上. 你也许会想这不是一个巧合，幕后一定还隐藏着些什么. 那么你就猜对了，163 的特殊性正是本章后面的部分所要揭示的，故事还得从古希腊讲起.

[57]

早期的数系

似乎是古希腊人最早建立起了算术的数学理论.爱奥尼亚学派(在约公元前600年由泰勒斯(Thales)建立)和毕达哥拉斯学派(由毕达哥拉斯在约50年以后创立)都发展了内容广泛的几何(特别是毕达哥拉斯学派)和算术理论.是希腊人首先认识到自然数(或计数数)1,2,3,…形成一个无穷的集合,并可在其中进行基本的加和乘的算术运算.虽然他们不承认负数是数,但他们懂得如何使用减号,如:

$$(7-2)\times(6-3)=(7\times 6)-(7\times 3)-(2\times 6)+(2\times 3).$$

他们的做法可能略有点像老式学堂用的顺口溜的意思:

"负负得正,正负得负;

无须证明,只管记住."①

然而希腊人不把 -5 这样的对象看做一个数是有相当理由的.对他们来说,数是与距离、面积和体积的量度紧密联系的.代数的法则通常用几何的术语进行思考,诸如将各种面积拼补粘合(见图4).

但是即使希腊人不需要负数,他们却肯定还需要分数或如数学家所称的有理数.(正)有理数是形如 a/b 的数,这里 a 和 b 都是自然数.因为 b 可以为1,所以有理数包括自然数(按第2章中的术语,自然数构成了有理数的一个子集).希腊人原来一直相信(正)有理数系对解决几何问题已经足够了,而到公元前6世纪的某一天,他们却惊恐地发现根本不是这么回事.特别是人们发现2的平方根不是有理数,这就意味着有理数不能用来准确量度诸如底和高都是1个单 [58] 位长的直角三角形的对角线(见图5).(为了能够量度所有的几何长度,就需要实数——我们很快会讲到更多有关实数的事.)这一发现实际上标志着希腊人终止了在算术方面的任何进步,他们从此把数

① 原文为"Minus times minus equals plus,The reason for this we need not discuss",直译为"负负得正,理由不用讨论".——译者注.

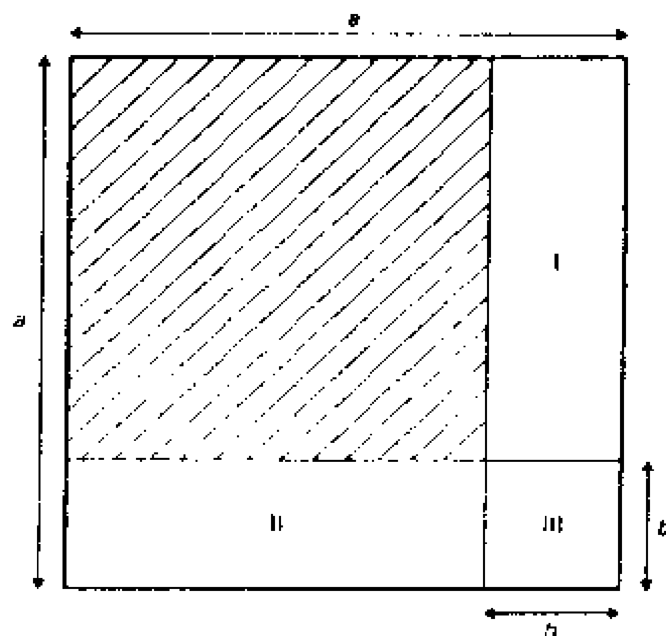


图4 希腊代数. 希腊人把熟知的代数等式,如$(a-b)^2=a^2-2ab+b^2$用纯几何的形式加以验证. 为了得到阴影面积,就要从整个面积(a^2)出发,减去由Ⅰ和Ⅲ组成的长方形(ab)及Ⅱ和Ⅲ组成的长方形(也是ab),再加上小正方形Ⅲ(b^2)以补偿多减去的重合部分. 这就给出了上面的等式.

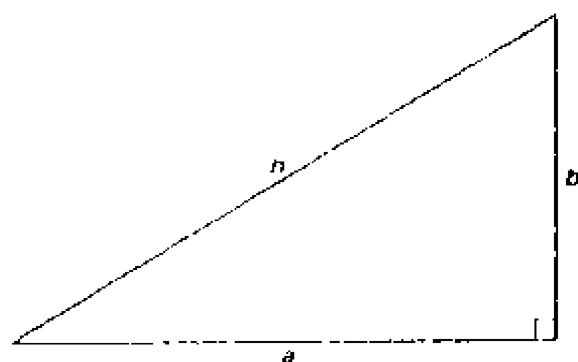


图5 毕达哥拉斯定理. 对任何直角边是a和b,斜边是h的直角三角形,公式

$$h^2=a^2+b^2$$

成立. 当$a=b=1$时,由公式得出$h=\sqrt{2}$,这是一个无理量,即它不能表示为两个整数的商.

[59]

学限定在几何构造的范围内.

负　　数

最早用到零和负数的自成系统的代数学是由7世纪的印度数学家所创立的.他们用正数和负数去处理包括借贷在内的财务问题.他们不仅最早使用了现代意义上的零,而且还写出过一些含有负数(在数字上加一点来表示)的方程,这是负号的早期表示法,并明确地提出了符号法则(正乘正是正,正乘负是负,负乘负是正).他们还认识到每个正数有两个平方根——一个是正的,另一个是负的.

印度人的这些早期工作并没有影响到14世纪到16世纪文艺复兴时期的欧洲数学家.后者沿袭着古希腊的传统,乐于使用负号但却不能像印度人那样接受负数.方程的负根被叫做“虚构的根”.

到了17世纪,一些数学家开始使用“负数”,但这种趋势受到了抵制,有时反对还来自数学界的名流.列涅·笛卡儿(René Descartes)讲过负根是“不真实的根”;帕斯卡(Blaise Pascal)同样认为比零小的数是不存在的;莱布尼茨(Gottfried Leibnitz)则同意负数会导致荒谬的结论,然而他也为负数辩护说:在进行计算时负数是有用的;欧拉接受了负数,但却相信它们比无穷(∞)要大,理由是因$\frac{a}{0}=\infty$,所以当我们用比0小的数去分a时,结果就要大于无穷.

只是到了18世纪,在代数中应用负数(用负号作标记)才最终流传开来,尽管那时许多数学家对负数还是感到不舒服,只要可能就不遗余力地避免使用负数.的确,只有接受了数的公理化思想以后(见第2章),负数才真正有了意义.这种说法同样适用于复数,但在我们讨论它们之前,应该说说“实数”. [60]

实　　数

虽然目前我们有关数系的讨论根据数的类型的不同分成几个部

分,但它们的历史的发展却不是这样泾渭分明的.负数、实数和复数的理论是在大致同一时期发展起来的.在三者当中,出色地处理实数是最杰出的成就(也是最难的),虽然(我们将看到)在复数理论中首先要假定实数的存在性,可是实数理论却是最后才建立起来的.

对所有实际的目标而言,有理数就已经够用了.在现实世界中(与数学的世界相对)的确只能用到这些数,问题的答案大多只用到几位小数.当然,有理数也有一些令人愉快的数学特性.两个有理数相加,相减、相乘、相除的结果仍旧是有理数,有理数的算术还满足第2章(第30~31页)中所列的整环的所有公理.数学家由此得出有理数构成了一个域.什么是域?是其中可做除法的整环,即满足第2章中第30~31页的七个公理和以下公理的一种结构:

(8) 对任意不为0的数 x,存在 y 使 $xy=1$.(乘法逆元素存在.)

容易验证(你不妨试一试):对任给的 x,由公理8所保证存在的 y 是唯一的.形式上把这唯一的数 y 写成 x^{-1},有时也写成$\frac{1}{x}$.公理8使除法可行,因为$\frac{a}{b}$就是 ab^{-1}.

[61] 简而言之,域是能进行通常的算术运算的一种结构.正如毕达哥拉斯学派所发现的那样,有理数域不能包含诸如方程 $x^2=2$ 的解.用有理数你可以求出达到任意精确度的解,如

$$1^2=1,\quad (1.4)^2=1.96,\quad (1.41)^2=1.9981,$$
$$(1.414)^2=1.999396,$$

等等,但没有一个有理数的平方准确地等于2;在另一方面,实数构成了一个包括有理数的域,这个域中元素丰富,足以解像上面那样的方程.这里给出精确的解的关键思想是由上例中的逐次逼近过程提供的.1,1.4,1.41,1.414,…给出了平方是2的那个数的越来越好的近似.要是能使用有无穷多位的十进小数,我们则能写出其平方恰为2的一个数:

$$1.414213\cdots(\text{直至无穷}).$$

可是我们显然写不出无穷多位的小数序列,那么实际上该怎么做呢?

答案是让数学为我们处理必要的极限概念，即是说必须用公理的方法来生成实数. 事实证明这是一个极其困难的工作，其内容远远超出了本书的范围. 为建立实数概念而需要的适当的公理体系在 19 世纪 70 年代最终完成，它是最伟大的数学成就之一.

实数包括所有的有理数（就像整数构成有理数的一个子集一样），同时也包括许多其他的数. 不是有理数的实数叫做无理数. 无理数的例子有 π，e 和$\sqrt{k}$（k 为不是完全平方的任意自然数）. [62]

复　数

16 世纪，欧洲数学家——特别是意大利人邦贝利（Rafaello Bombelli）——开始认识到，在解代数方程时假设负数可以开平方根常常是有用的. 考虑到那时的学术"气候"，你就会理解为什么这样的数被称作是"想象的数"，而现在数学家把所有的数都看做是"想象的"概念，负实数的平方根并不比其他的数特别. 不管怎么说，现在仍习惯于把负数的平方根叫做虚的（想象的）数，于是这里"虚的"（想象的）这个词就有了一种特殊的、技术性的意义.

事实上，为得出负实数的平方根，只需假定方程 $x^2 + 1 = 0$ 的解存在. 如果 i 代表这一方程的解（$i^2 = -1$），则对任意正实数 a，负数 $-a$ 的平方根将是$(\sqrt{a})i$（实际上有两个平方根$(\sqrt{a})i$ 和 $-(\sqrt{a})i$. 同样，方程 $x^2 + 1 = 0$ 也有两个解，i 和 $-i$）. 形如 ai 的数叫做虚数，这里 a 是实数（字母 i 的这种用法最早由欧拉开始）.

复数是形如 $a + bi$ 的数，其中 a 和 b 是实数. 这里的加号并不是一般的加的意思（实数、虚数怎么相加?），而是把复数的实部 a 与虚部 bi 分开. 注意，如果 $b = 0$，则 $a + bi = a$，所以实数构成了复数的一个子集. 同样地，如果 $a = 0$，则 $a + bi = bi$，所以虚数也构成了复数的一个子集.

至此，就算你容忍了 $i = \sqrt{-1}$，你也可能会想，把形如 $a + bi$ 的东西叫做数有什么道理呢? 不过请记住，复数是什么模样并不重要，

重要的是它们如何运作. 如果复数具有一种有效实用的算术(不管是 [63] 在数学内部或是更广的范围内), 有可能构成一个域, 则它们也应有被叫做"数"的权力. 那么复数的算术是什么样的呢? 其规则如下. (对大多数人来说, 这是他们所遇到的第一个公理化观点下的数系. 在对整数, 有理数, 或实数中的任何一种进行公理化之前, 大多数人就已经对它们相当熟悉了.)

两个复数相加的规则很简单: 把实部相加, 再把虚部加在一起. 于是,

$$(a+bi)+(c+di)=(a+c)+(b+d)i.$$

举例来说:

$$(2+3i)+(7+1i)=9+4i,$$
$$(-3+4i)+(4-2i)=1+2i.$$

复数的乘法略微复杂些, 思路是使用一般的两个带括号的和相乘的代数法则, 然后让 $i^2=-1$. 这样,

$$\begin{aligned}(a+bi)(c+di)&=ac+adi+bci+bdi^2\\&=ac+adi+bci-bd\\&=(ac-bd)+(ad+bc)i.\end{aligned}$$

例如:

$$\begin{aligned}(2+3i)(5+7i)&=10+14i+15i+21i^2\\&=10+14i+15i-21\\&=-11+29i.\end{aligned}$$

令人惊奇的也许是复数可做除法, 规则是:

[64]

$$\frac{a+bi}{c+di}=\frac{ac+bd}{c^2+d^2}+\frac{bc-ad}{c^2+d^2}i.$$

例如:

$$\begin{aligned}\frac{3+5i}{1+2i}&=\frac{3\times1+5\times2}{1+4}+\frac{5\times1-3\times2}{1+4}i\\&=\frac{3+10}{5}+\frac{5-6}{5}i\\&=\frac{13}{5}-\frac{1}{5}i.\end{aligned}$$

事实上，复数构成了一个域.（作为练习你可以验证，由上面定义的加法和乘法确实能得出满足域公理的算术.）所以，虽然你会感到复数概念有些奇怪，可它却提供了一种"一般"类型的算术. 而且，你能从中得到其他数系所没有的额外收获：在复数域中，每个多项式方程都是可解的！即若 $a_0, a_1, \cdots, a_{n-1}, a_n$ 是复数，则存在复数 x，它为方程 $a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0$ 的解. 当然实数就没有这样的性质，方程 $x^2+1=0$ 已证实了这一点.

刚才提到的结果叫做代数基本定理. 1692 年由吉拉尔（Girard）首先提出，达朗贝尔（d'Alembert）（1746）和欧拉（1749）都曾证明过它，但证明不完善. 第一个完全正确的证明是由高斯在 1799 年他的博士论文中给出的. 高斯对这一结果印象很深，后来又给出了三个（完全不同的）证明.

代数基本定理只是使复数系成为这样一个"好"的数系的原因之一. 另一个重要的原因是复数域支持了强有力的微分学的发展，导致了多产的复变函数理论.（这一理论将在第 9 章中谈及.）

复数理论不仅在数学上是吸引人的，实际上也是极其有用的. 复数的第一个精彩的科学应用是由斯泰因米茨（Charles Steinmetz）做出的，他发现复数在涉及交流电的高效率计算中发挥了实质性的作用. 今天没有哪个电气工程师可以离开复数，搞空气动力学和流体力学的人也是这样. 爱因斯坦的相对论中也用到了复数：三个空间维数看做是实数，而时间维数看做是虚数；就是在量子力学中，物理学家也必须和复数打交道. [65]

尽管复数构成了一个域，尽管它们非常有用，尽管所有的数系其实都是纯抽象的、"想象的"结构，许多人还是对复数感到不自在. 这主要（完全?）是一个习惯问题. 比如，在分析学家的"显微镜"下，实数可能是非常复杂的数学对象. 而两面趋向无穷，0 在中间的实直线却是一幅令人非常舒服的简单图形（见图 6）.

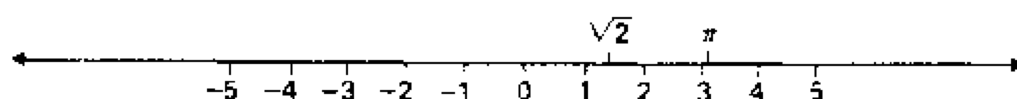


图 6 实直线. 实数公理保证了此线为连续的,中间没有"空隙",甚至没有只能容下一个点的无限小的空隙(在这种意义下,"有理直线"在$\sqrt{2}$应在的地方有一"空隙").

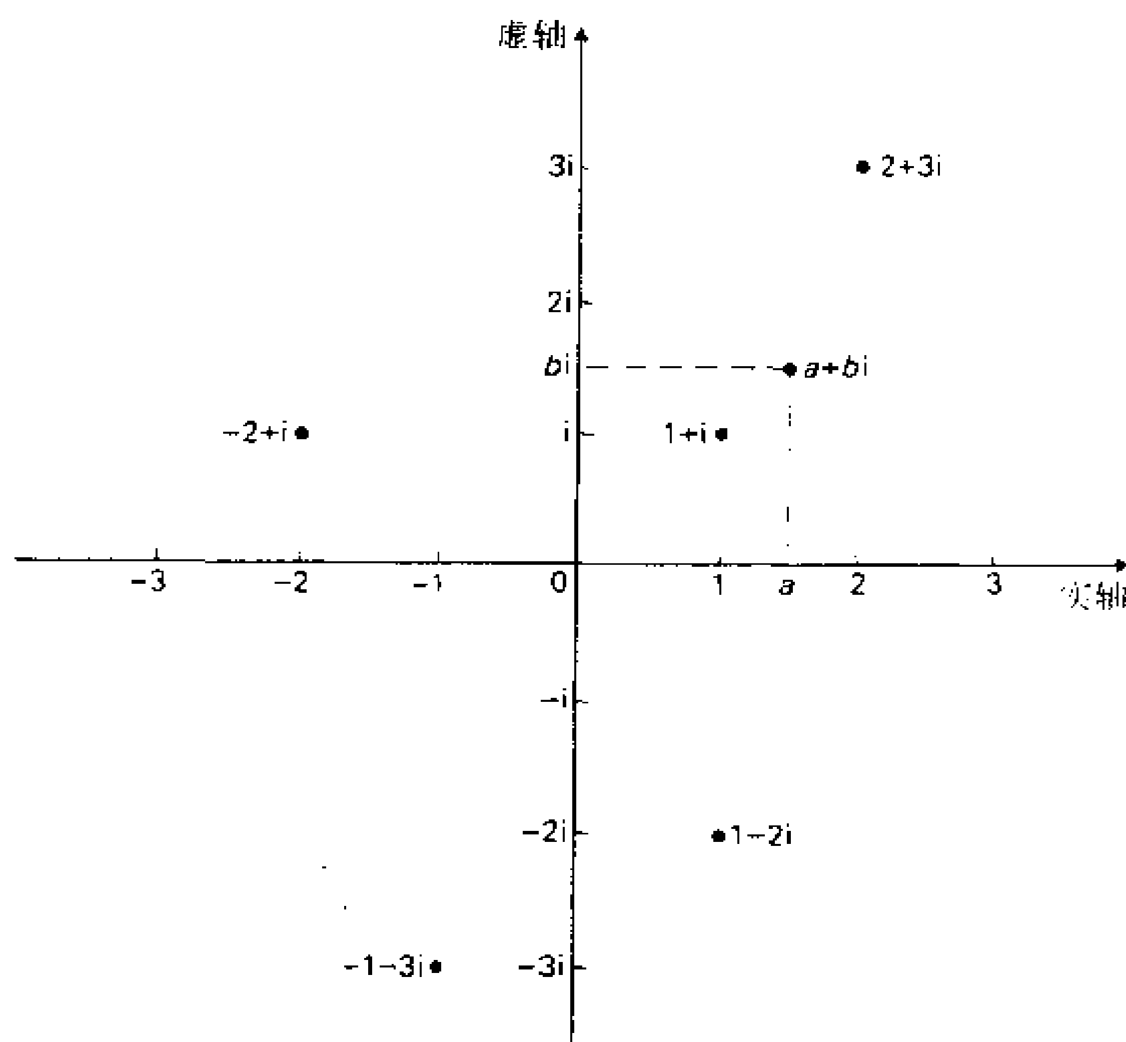


图 7 复平面. 复数 $a+bi$ 由数对(a,b)确定的点识别. 实数在水平轴上,纯虚数在纵轴上.

[66]

复数有同样令人舒服的图，这是个好消息．就像实数可以看做是实线上的点一样，复数可以认为是二维平面上的点(见图 7)．第一个把复数这样形象化的是自学成材的挪威测量员维赛尔(Caspar Wessel)，1797 年他报告了他的想法．瑞士藏书家阿甘得(Jean－Robert Argand)和高斯与他所见略同．阿甘得在 1806 年出版了一本书讨论这个问题，事实证明此书产生了极大影响，复平面——二维平面用来表示复数时的称呼——有时也叫做阿甘得图．

四 元 数

复数可以表为复平面上的点，在这一思想启发下，爱尔兰数学家哈密顿(William Rowan Hamilton，1805～1865)利用实数坐标给复数一种代数的(即本质上是公理化的)解释．他接着尝试把复平面推广到三维的可能性．事实证明这是行不通的，但哈密顿发现了在四维情况下可以推广出一种可叫做“超复数”的系统．

[67]

哈密顿把他发现的新数叫做*四元数*，这项发现来之不易，经过几年的思考才最终取得关键性的突破．正像在数学研究中经常发生的那样，重要的思想并不是坐在书案旁得来的．1843 年的一个黄昏，他和妻子沿着都柏林的皇家运河散步，突然他意识到假如他舍弃乘法交换律，那么其余的一切就能正常运转了．(即他可以得到另外一种可接受的数系．)他的想法使他兴奋不已，他停在布罗罕桥上把基本的公式刻在一块石头上．(原来的刻痕早已被日月风雨侵蚀殆尽，但那座桥上现在装嵌了一块纪念牌以记载那个伟大的事件．)

简单地说，四元数是一种形如

$$a+bi+cj+dk$$

的数，这里 a,b,c,d 是实数，i,j,k 是满足等式 $i^2=j^2=k^2=-1$ 的“虚”数，哈密顿刻在桥上的关键的等式是：

$$ij = k, \quad jk = i, \quad ki = j,$$
$$ji = -k, \quad kj = -i, \quad ik = -j.$$

用这些规则，任意两个四元数可以相乘——即得到第三个四元数.（加法跟复数的一样，只要简单地将对应项相加即可.）按一般的代数法则所形成的数系满足除了乘法交换律之外的整环的所有公理.

在现代物理学中发现四元数有相当大的用处，更奇怪的一种数系——八元数亦然，后者是一种八维的数系，其中不仅交换律不成立，就连乘法结合律也不成立.现在是我们回到自然数上面，看看高[68]斯在数论上的工作的时候了.

高 斯 整 数

1796年，高斯证明了数论中的一个深奥的定理，它叫做二次互反律，是用来解像

$$x^2 \bmod 7 = 3$$

这样的方程的.这种方程的一般形式为

$$x^2 \bmod p = q,$$

其中 p 和 q 是素数.在试图把他的理论推广到包括高次方程的情形（$x^3 \bmod p = q$ 等）并最终得以成功之际，高斯发现如果使用形如 $a + bi$（其中 a 与 b 是整数，$i = \sqrt{-1}$）的数，而不是仅仅用到整数，他所进行的计算就会简单一些.现在，这种"复整数"叫做高斯整数.它们在分解因子时特别有用.正像普通整数有因子分解：

$$a^2 - b^2 = (a + b)(a - b)$$

一样，高斯整数导致

$$a^2 + b^2 = (a + bi)(a - bi).$$

从直觉上看高斯整数似乎应占据复数域中像普通整数在实数域中那样的位置.但高斯整数究竟跟整数有哪些相像呢?

在第1章中已搞清楚，有关整数的最重要的事实包含在算术基本定理中：即每个整数可唯一地表示为一些素数（不能进一步分解的

数)的积，最多相差-1. 高斯证明了在高斯整数中有些数是“素”的 [69] (即不能被分解). 利用这样的“素数”可以对高斯整数得到类似于算术基本定理的唯一因子分解定理.(这里的“素数”并不是指形如 $a+bi$，而 a 与 b 都是素数的数. 高斯素数定义为那些不能分解为其他高斯整数的积的高斯整数. 正因如此，数学家们经常称它们为不可约的.)

类 数 问 题

实践证明，除了互反律，高斯整数还在别的问题上很有用处，其中出名的是它们跟费马大定理的联系，这部分内容在第 8 章中有更多的介绍. 高斯整数被证明如此有用，以至探讨别的跟它类似的数系也变得很有意义了，这正是高斯所做的. 在可能探讨的各种数系当中，最富成果的是形如 $a+b\sqrt{-d}$的数所构成的数系，这里 d 是某个不为 1 的正整数.

现在，还有一件能引起小小的惊奇的事情. 为了得到一个合理的数系(即与普通的整数有某种相像的数系)，当取 d 满足 $d \bmod 4=3$ 时，你就得让 a 和 b 既能取为整数又可取为半整数. 如

$$\frac{1}{2}+2\sqrt{-3},\quad \frac{3}{2}+\frac{5}{2}\sqrt{-3}$$

是取 $d=3$ 所得的数系中的数(如果 $d \bmod 4 \neq 3$，那么跟高斯整数一样，a 和 b 就只取整数).

作了上面的小小的修正以后，你会问哪一个 d 值会使你得到一种合理的“数论”. 特别地，哪个 d 值将使你可以得到唯一因子分解定理？ $d=1$ 时(高斯整数)，有此定理；$d=2$ 和 $d=3$ 时也能得到这个定理，但 $d=5$ 时就不能，因为在这个数系中 6 这个数有两种形式的因子分解(分解至不可约的情形)：

$$6=2\times 3, 6=(1+\sqrt{-5})\times(1-\sqrt{-5}).$$ [70]

在高斯的时代，知道有 9 个 d 值使 $a+b\sqrt{-d}$所产生的数系有唯一

因子分解定理(a,b 如上面所指出的那样取值).这9个值是:

$$d=1,2,3,7,11,19,43,67,163.$$

还有别的值吗?虽然高斯和其他人在随后的几十年里几经努力,可还是没有再找到.下一个结果是1934年海布龙(Heilbronn)和林福(Linfoot)作出的,他们说至多还可存在一个其他的 d 值,如果它存在,那将是一个天文数字.但是有没有第10个这样的 d 值呢?

1952年,有一个人知道了不存在那个 d 值.那一年,一位把数学当作业余爱好的退休的瑞士科学家希格内尔(Kurt Heegner)发表了他的证明,声称第10个 d 值不存在,但没人相信他.(他的文章很难读,虽然如此,……)世界上其余的人在等待了15年之后才知道了这一真理.麻省理工学院的斯塔克(Harold Stark)和剑桥大学的阿兰·贝克(Alan Baker)独立地(用不同的方法)证明了第10个 d 值不存在,这一次数学界信服了.在这一发现的鼓舞下斯塔克和贝克又重新检查了希格内尔的工作,结果惊奇地发现希格内尔的证明本质上是正确的.被忽视的瑞士人其实是对的!

现在你能知道为什么163是这么特殊了吧?在本章开始提到它引起了那些奇怪的结果.163是使 $a+b\sqrt{-d}$ 这样的数系有唯一因子分解的最大的 d 值.(不幸的是这里不能进一步指出究竟 $d=163$ 时的唯一因子分解特性与早先谈过的它的性质有什么关系.那纯粹是职业数学家的事!)

在搞清了那些使唯一因子分解定理成立的数系 $a+b\sqrt{-d}$ 之后,我们对那些没有唯一因子分解性的数系又该说些什么呢?高斯又一次扮演了领头羊的角色.他把从 d 值得来的每个数系与某个自然数 $h(d)$ 连系起来,这个自然数叫做那个数系的类数(class number).这个类数可以确定唯一因子分解定理失效的界限.如果类数是1(高斯所列出的那些 d 值都属于这一情形),则唯一因子分解定理成立,如果 $h(d)=2$(比如 $d=5,6,10,13$),则唯一因子分解定理失效.类数为3时($d=23,31,59$ 等),则定理失效的程度稍有增加,类数为4时($d=14,17,21$)失效的程度更高,以后的情形大致如此.类

[71]

数越大，在该数系中把数分解成(该系中的)素因子的方法越多.

在《算术研究》(高斯的不朽著作，我们在第1章中提过)第303节中，高斯描述了一些类数的粗略的计算，并注意到对每个类数k，似乎有一个满足$h(d)=k$的最大的d值.使$h(d)=1$的最大的d值是163，使$h(d)=2$的最大d值是427，使$h(d)=3$的最大d值是907.虽然高斯猜想情形一定是这样的，但他既不能完全肯定刚才提到的值就是那些最大的，也不能证明总有最大的d值存在.

类数问题(要假定高斯的猜想成立，以便使下面的讨论有意义)就是要对每个类数k确定使$h(d)=k$的最大的d值.(希格内尔1952年的结果解决了$h=1$时的类数问题.)

从高斯时代开始直到这个世纪，类数问题实际上没有任何进展.1916年，海克(Hecke)证明了如果一个叫做广义黎曼猜想的更复杂的陈述正确的话，那么每个类数只与有限多个d值相关的高斯猜想就成立.可是因为没有人知道广义黎曼猜想对还是不对(现在仍是如此)，海克的结果没说出很多信息，或至少对素数问题本身没说出什么.1934年，在多灵(Deuring)和莫戴尔(Mordell)的近期工作的基础上，海布龙证明了在广义黎曼猜想错误的假设下高斯猜想成立.因为所论及的黎曼猜想非对即错——尽管我们没有去决定(或如第2章中所说的，心里有了一种结论，但不能肯定)哪一个正确——海克和海布龙的结果加在一起最终证明了高斯猜想.

高斯猜想一经证明，解决类数问题本身的道路即已扫清.首先已经有希格内尔1952年的结果证明了$h=1$的情况，接着在1967年，在重新证明了$h=1$的情况后贝克与斯塔克又揭示了$h=2$的情况.但已有的方法都不能解决其他情况下的问题.

重大突破发生在1975年，位于奥斯汀的得克萨斯大学的哥德菲尔德(Dorian Goldfeld)得到了部分解答.经过解析数论方面的冗长而又艰难的论证，哥德菲尔德证明了假如能得到某一(更准确地说是复的)数学对象的话，则类数问题就可以完全解决.他所要求的数学对 [72]

象是具有某种形状的几何曲线[①],它还应具有一些不同寻常的特殊性质.找到所需形状的曲线并不难,问题是要找到具有某种特殊性质的这类曲线.哥德菲尔德努力想找出一个,可却失败了,就像以前研究这一问题的那些人一样.

直到1983年,蔡格尔(Zagier)和格罗斯(Gross)才在他们的研究中取得成功.他们的关键性的思想是:寻找在曲线上的一些特殊点,为了纪念被长期忽视的希格内尔,他们把这些特殊点叫做希格内尔点.那时的证明中出现了一个庞大的方程,光是计算方程两边的项就用了100页.然后他们还得把项成对地列出以验证方程的正确.证明的这一部分虽然很长,但在数学家看来是"明白易懂"的.真正令人吃惊的是一条简单的曲线居然控制着无穷多族数系的性状.

经过183年的历程,高斯的类数问题最终得到了解决.

阅读文献

有关数系的历史演进情况可参看 Graham Flegg 的 Numbers:Their History and Meaning(André Deutsch,1983).

Ivan Niven 的 Numbers: Rational and Irrational(Random House,1961)简单地描述了数系,更全面的介绍可以在 Claude Burrill 的 Foundations of Real Numbers (McGraw-Hill,1967)以及 Leon Cohen 与 Gertrude Ehrlich 合著的 The Structure of the Real Number System(Van Nostrand,1963)中找到.

对在类数问题中出现的各类数系的清晰描述可见 Harold Stark 的 An Introduction to Number Theory(Markham,Chicago,1970)一书的第8章.内容更深一点的书可参见 I.N.Stewart 和 D.O.Tall 的 Algebraic Number Theory(Chapman and [73] Hall,1979).

Don Zagier 的文章 *L* - series of elliptic curves, the Birch - Swinnerton - Dyer conjecture, and the class number problem of Gauss 是叙述类数问题的最终解决的

① 应特别指出,这种曲线应具有如下形式的方程:

$$y^2 = ax^3 + bx^2 + cx + d.$$

这种曲线叫做椭圆曲线,它们除了在此提到的应用外,在数论中还有许多应用.——原注.

一篇高水平论文，刊登在 Notices of the American Mathematical Society, Volume 31, Number7(November 1984), Issue 237, pp. 739 - 43 上. 这篇文章也为愿意进一步深入了解的人提供了更多(高层次的)文献. 但应指出，此著作非常高深，大多数职业数学家都会感到难以理解其中的证明.

最后，如果想亲自一读高斯的 Disquisitiones Arithmeticae，耶鲁大学于 1966 年出过一个英译本(原著于 1801 年在莱比锡出版). [74]

(李家宏译)

第4章 混沌之美

数学的美

罗素(Bertrand Russell)在1918年出版的《神秘主义与逻辑》一书中写道:

"公正而论,数学不仅拥有真理,而且拥有至高无上的美——一种冷峻严肃的美,就像是一尊雕塑."

另一位著名的英国数学家哈代,在《一个数学家的辩白》(1940年)一书中则写道:

"数学家的造型与画家或诗人的造型一样,必需美;概念也像色彩或语言一样,必须和谐一致.美是首要标准,不美的数学在世界上是找不到永久地位的.……数学的美很难定义,但它却像任何形式的美一样地真实——我们很可能不知道什么才算是美的诗,但这丝毫也不妨碍我们在朗读一首诗时去欣赏它的美."

两位作者在这里谈到的是一种形式高度抽象的美,一种为一切职业数学家所共识而对大多数人来说却未曾体验甚至难以想象的内在美.这是一种逻辑形式、结构与证明的美,一种只有经过长期艰苦[75]探索之后才能领略的美.

至少到本世纪80年代初为止情况一直如此.后来由于电子计算机特别是图象显示系统的发展,某些新兴的数学使一切发生了变化.混沌动力学就是计算机开辟的一个新的数学领域,它还有其他不同的名称.虽然这门学科所涉及的某些数学知识与数学家的其他任何

艺术同样困难、同样抽象，但它所获得的结果可在计算机屏幕上显示出来，使所有的人，无论是数学家还是门外汉都有目共睹．在德国歌德学院组织的一个展览会上，计算机图象显示的硬片成为中心内容，这个展览会在 1985 年开始周游世界各地，在大学数学系和公共艺术陈列馆受到同样的欢迎．电影工业也很快意识到这门新数学的潜力．[76]来自复动力学（同一领域的不同名称）的日益增多的概念正被应用于科幻影片的图象制作．

图 8 只是这个新领域中各种常见结构的大量图示的一例．（这些图象有许多可制成彩色以突出黑白图中看不到的形状．）尽管令人感

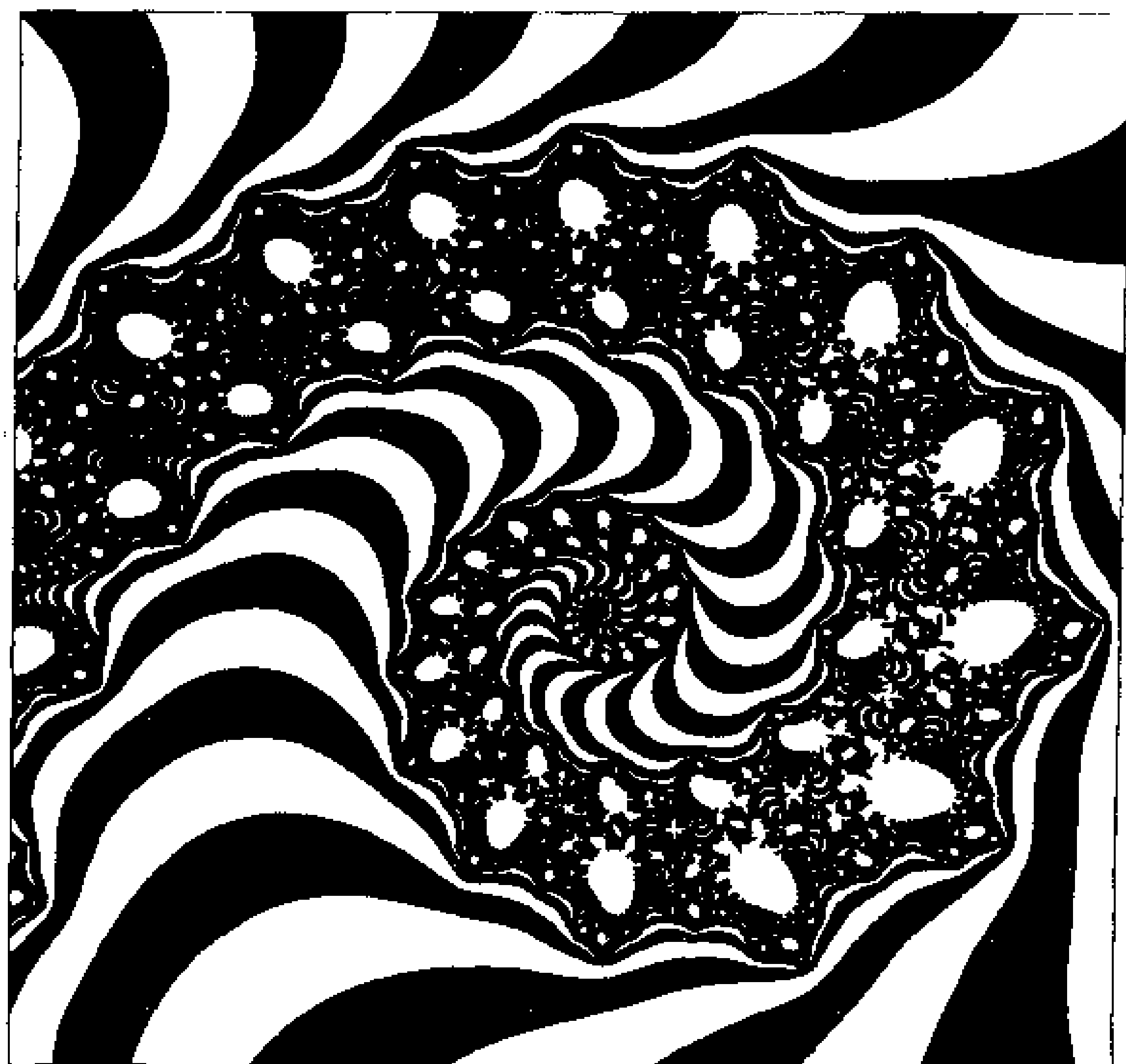


图 8 分形艺术——蒙德尔布罗世界一瞥．

到意外，但图8的复杂性却是某种简单数学的结果（虽然详细的分析也会涉及非常高等的方法）．本章将介绍这种数学．

英国海岸线有多长？

这问题正是1967年《科学》杂志发表的一篇划时代文章的标题，作者蒙德尔布罗（Benoit Mandelbrot）是一位杰出的法国数学家，当时正在纽约约克敦高地IBM公司的沃特生研究中心工作．问题乍看似乎很简单．你也许会想：利用地图或航空测量就能获得令人满意的答案．麻烦的是，无论你做得多么认真细致，都不可能得到准确的答数．而且有充分的理由认为：根本就没有准确的答数！这是蒙德尔布罗得出的惊人结论，他的推理如下．

假定你乘一架喷气式飞机在10000米高空沿海岸线作飞行测量，同时不断拍摄海岸照片，然后选取适当的比例尺并计算你拍到的大量照片所描示的整个长度．这样得到的答数是否精确呢？否！从10000米高空你不可能区别许多的小海湾和小海岬（假设你使用的是一架性能良好的普通型相机）．如果你改乘一架小飞机在500米高处重复这种测量，将会看清许多原来看不见的细部，而使你的答数大大增加．在第一次拍摄的照片中显得光滑的海岸线，现在被发现包含着无数小海湾和小海岬．

[77]

现在假设你降落地面，用量规来测量海岸线长度，间隔比如取1米，那么那些在空中看不清的岸儿将使答数变得更大．如果你改取间隔为10厘米，结果继续增大，如此等等．每一次，度量越精密，海岸线就显露出越多的细节，而你获得的答数也就变得越大．很快你就会去测量石子、沙粒，然后是分子等等．在所有情形，你得到的答数都将不断增大．

当然在物理世界，这种越来越精细的测量过程必然会有终结．就人的限制而言，你可能会在使用1米间隔的量规后就停止测量，而物理学家可能会认为这种测量过程必将在原子层次上达到一个理论的极限．但从数学家理想化的观点看，这种越来越精细的测量过程则可

以无限继续下去. 这意味着相应的测量结果将无限地增大. 也就是说,所谓海岸线的长度并没有确切的数学定义,而仅仅是任意的选择——这种选择甚至不能看作是某个"真实"答数的近似.

冯·柯克(H.von Koch)在1904年首先考虑过的一种几何图形,为蒙德尔布罗的不可捉摸的海岸线问题提供了理想的数学模型.我们把这种几何图形称为柯克岛.图9(i)表示从一艘外空火箭上看到的柯克岛.从这样的距离看,其形状仿佛就是一个等边三角形.当火箭飞近地球时,逐渐看清了这个等边三角形的每条边上,还包含一个海岬,它们形成一个小的等边三角形,位于每边中央三分之一处(图9(ii)).如果图9(i)的周长为3,那么图9(ii)的周长将是 $3\times\left(\frac{4}{3}\right)=$ [78] 4.当火箭进一步靠近地球,你会发现原先看到的十二条小边每一条同样包含有一个位于中央三分之一处的等边三角形海岬(图9(iii)).现在图形的周长变成 $3\times\left(\frac{4}{3}\right)\times\left(\frac{4}{3}\right)$.图10是靠得更近时所看到的岛的样子,显露出更多层次的细节,使我们对柯克岛的"真正"(?)形状能有某种印象.

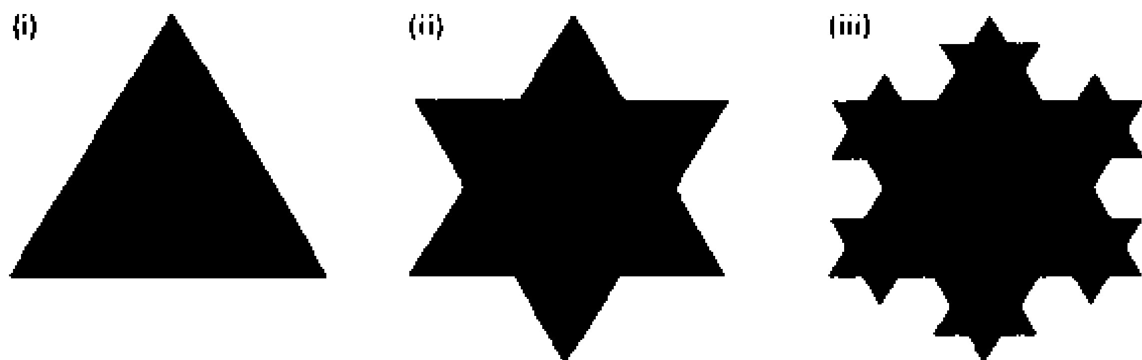


图9 柯克岛的构造.

对数学家来说,柯克岛有一个很好的特征,即相继层次的细节所显示的规则性.每一步,海岸线每条线段的中央三分之一都被两条小线段取代,其中每一条都等于原线段的三分之一,如图11所示. [79]

通过对图9和图10的考察,你也许会推测柯克岛具有一个(数学上)确定的形状.而在人眼所能区别的范围内图10是它的很好的

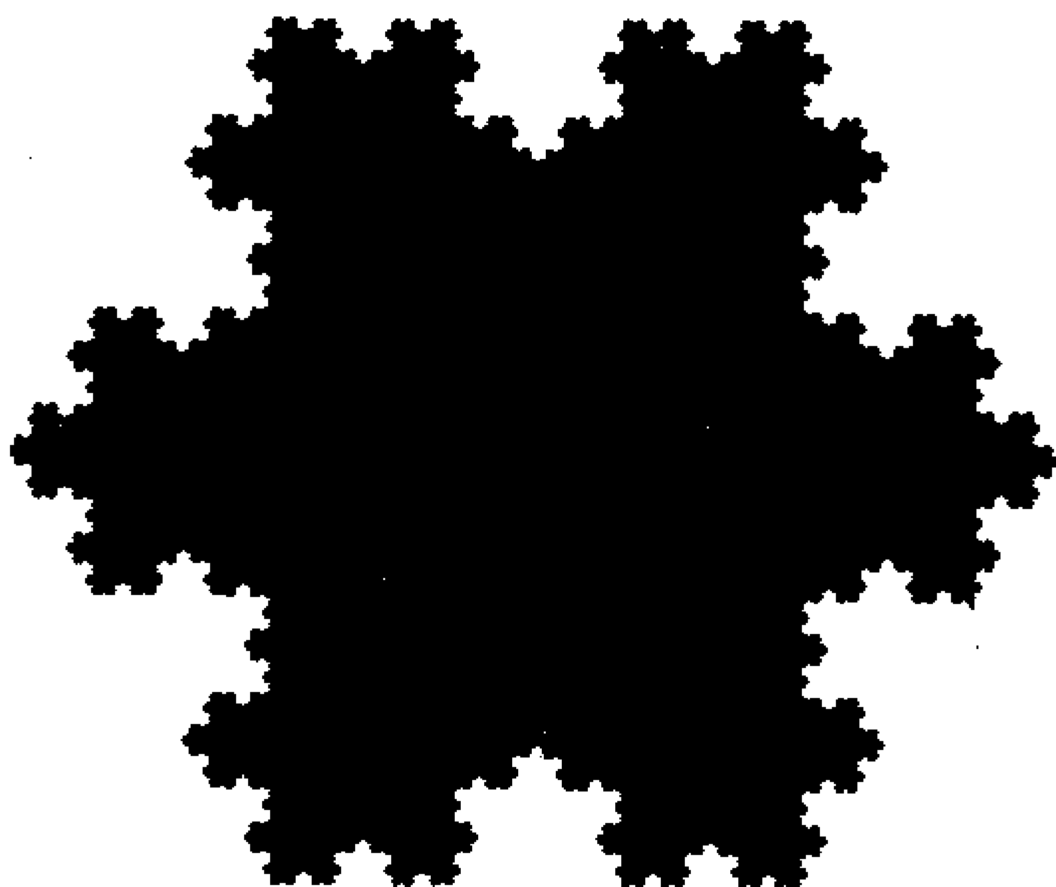


图 10 柯克岛成形.

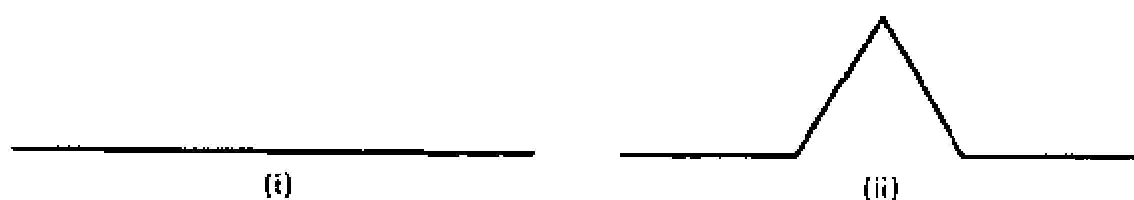


图 11 柯克海岸线的产生.

近似.在数学上,柯克岛的精确海岸线是这样一条"曲线",它是一个无限逼近序列的极限,图 9 给出了该序列的前三个图形.从数学上看,这条极限曲线是完全确定的,像任何其他曲线一样由无限多个点

组成，这些点排列在一起而形成“线”. 这种极限曲线的过程类似于将数$\frac{1}{3}$看成无限十进小数序列

$$0.3, 0.33, 0.333, 0.3333, 0.33333, \cdots$$

的极限的过程.

因为柯克岛是一个数学上确定的平面区域，它将有一个确定的面积. 这面积的具体数值当然取决于所使用的测量单位，但它肯定是有限的.（它可以作为一个数列的极限而计算出来，颇似上述例子中的数$\frac{1}{3}$. 事实上，它恰好等于图 9（i）中三角形面积的 1.6 倍）. 那么，围绕这块有限面积的海岸线有多长呢？柯克过程的每一步都使“海岸线”的长度扩大$\frac{4}{3}$倍，当柯克曲线（作为极限海岸线的称呼）被达到时，扩大因子$\frac{4}{3}$出现了无限次，因此柯克曲线的长度将是无限大数.

一块有限的面积何以会有无限长的边界呢？图 9 和图 10 本身已提供了回答. 边界曲线沿着整个长度不断折曲，对于向终极曲线的每一步有限逼近来说，只要采用适当的比例（放大率），这种折曲就可以被完全画出来，然而真正的柯克曲线却有无限多折曲，于是某种非常奇怪的事情发生了，这就是新维数的出现. [80]

新　维　数

我们在几何中通常遇到的曲线都是一维的：例如一个被限制在直线或圆上活动的生物只能沿一个方向旅行（假若将向后运动看作是负的向前运动）. 通常的几何曲面如平面或球面是二维的：上面有两个独立的旅行方向，一般用向前/向后与向左/向右表示. 立体的物体是三维的，允许三个方向的运动. 火车提供了一维运动的例子，船舶可以在海面上作二维运动，而飞机则能作三维运动.

就人类的经验而言，我们生活的宇宙是三维空间（虽然相对论将时间看作是“第四维”，另外某些现代物理理论认为宇宙有 11 维，即

普通物理上所知的3维再加上8种被宣告为自然基本力如引力、磁力等的维度).但对数学家来说,三"维"并没有什么特殊的地方,可以如法炮制考虑四维或更高维的"空间".这类高维空间虽然不能按传统几何去理解,但它们却能有各种实际的应用.(一个例子是叫"线性规划"的学科,将在第11章中讨论.)不过请注意,这里所说的"高维",仍然是整数维.

所有这一切是否适用于柯克曲线呢?作为一条曲线(数学意义的曲线,虽然它有无限多个不能画出的折曲),你可能会以为它是一维的,其实不然.尽管柯克曲线通过上述方法得到的每一条近似曲线都是一维图形,但极限曲线却不是.由于其方向改变了无限次,我们进入了一个陌生的世界——实际上,"方向"这个词已不复适用.因此我们不能指望通过所谓"旅行方向"来确定柯克曲线的维度.我们必须寻找其他的、与方向无关的途径来建立维数的概念.

[81] 合适的做法是采取符合柯克曲线特性的途径.关键的性质是所谓自相似:部分与整体相似(只是缩小了比例).

假设有一个 D-维图形,将它分成 N 个与整体相似的部分.那么整体图形与每个部分之间的相似比 r(即整体比部分的放大因子)将由下式确定:

$$r = \sqrt[D]{N}.$$

(因为是 N 维图形,r 的值应当按单个维度计算,所以取 N 的 D 次根是必要的.)

例如,假如我们取一根直线,将它等分成 N 段(图12),那么每一段恰好等于整体长的 $\frac{1}{N}$.因此相似比为 N,这恰好就是当 $D=1$ 时由上述公式所得到的数值.

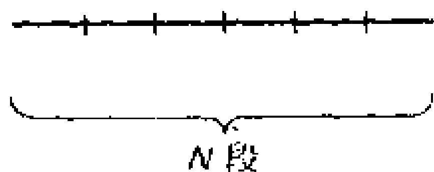


图12 直线段的自相似.

再如取一个矩形($D=2$),通过垂直与水平 k 等分而将它分成 N 块(图13),那么情形如何呢?这时整个图形恰好被分成 $N=k^2$ 个与整体相似的全等小矩形,而整体对于任何一小部分的(线性)比 r 由

[82]

$r = \sqrt[D]{N} = \sqrt[2]{N} = \sqrt[2]{k^2} = k$ 给出，这恰好又是你所期望的数值.

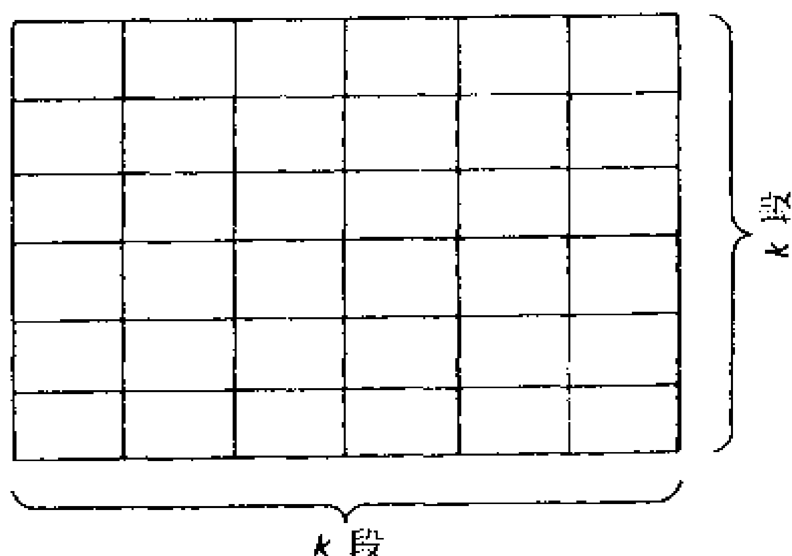


图 13 矩形的自相似.

在这两种场合，我们似乎一直在兜圈子，但这是因为我们处理的是司空见惯而又毫无疑义的情形. 假如我们用同样的方法去分析柯克曲线，就会得出完全出乎意料的结论. 对柯克曲线我们不知道它的 D，但 N 和 r 的值却很容易确定. 为此只需考察一下产生该曲线的复制过程. 我们首先来看海岸线的一段（图 11(i)）——任何一段都行，因为全都同样. 在复制过程中（图 11(ii)）每一段都被 4 条线段代替（因此 $N = 4$），每小段都等于原线段的三分之一（因此 $r = 3$）. 这对海岸线的每一段都成立，所以对整个海岸线亦必成立，于是根据前面确定的公式有

$$3 = \sqrt[D]{4}.$$

那么 D 等于多少呢？当然不是整数. 唯一的计算方法是利用对数. 若在上述方程式两边取对数，将得到

$$\log 3 = D^{-1}\log 4.$$ [①]

借助对数表（或利用能算对数的计算器），就可以算出 D 值，若精确

① 原文误作 $\log 3 = D\log 4$. ——译者注.

到小数后4位,则有

$$D=1.2618.$$

这样,柯克曲线是一个具有分数维的数学实体.

不只是柯克曲线可以有分数维.利用自复制过程还可以构造出同样稀奇古怪的"曲面"和"立体".例如从一个正方体开始,相继挖除"中间"部分,最终(即重复无限多次后)将得到一个叫谢尔宾斯基(Sierpinski)海绵($D=2.7268$)的立体,其构造如图14所示.这个不可思议的立体被一张无限"曲面"包围且体积为零.这海绵体的每一块表面被称为谢尔宾斯基毛毯,被一条无限边界线围绕而面积为零.[83] 谢尔宾斯基毛毯的维数 $D=1.2618$,与柯克曲线相同.通过对图14的考察并利用公式

$$r=\sqrt[D]{N},$$

或两边取对数

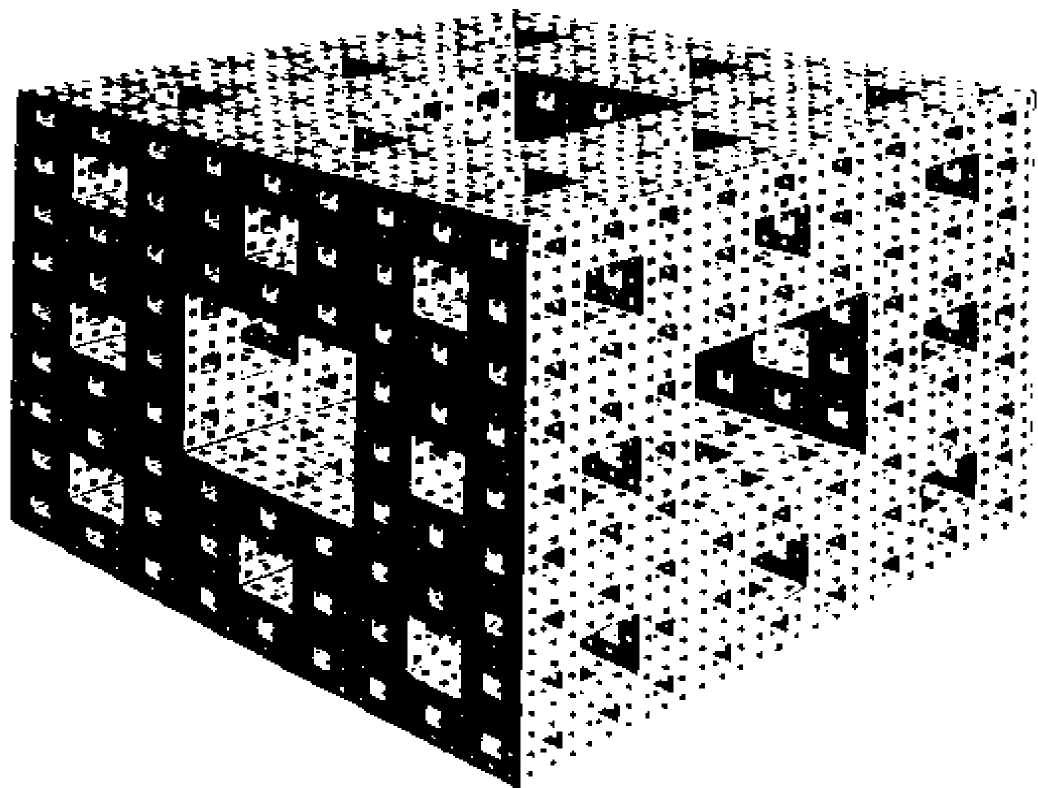


图 14 谢尔宾斯基海绵成形.

$$D = \log N / \log r,$$

就可以验证与谢尔宾斯基海绵有关的两个 D 值.

蒙德尔布罗 1977 年把具有分数维的图形称之为分形(Fractals),分形几何就是以这类图形为对象的数学分支.

本章其余部分仍与分形有关,但却讨论与柯克曲线和谢尔宾斯基海绵类型不同的分形.这两种分形具有高度的规则性,自复制过程的每一步都相同.如果将图形的某个特定部分放大作更细致的观察,也不会发现什么异常——步步相同,直至无穷.1980 年以后,计算机被用于考察自复制过程不断改变(虽然如后面将要说明的那样,这样的过程仍被称为"自复制")的分维图形.对于这样的图形,图象放大将会显示出完全出乎意料的结果,图 8 只是一个例子.这样的分形既是数学研究的对象,又是实验(以计算机为工具)的对象,它们将研究者引进了一个新奇诱人和美丽无比的世界,其发现也像其他许多新世界一样,部分地是缘于偶然的机遇. [84]

发现新世界

到 1978 年,蒙德尔布罗关于分形的工作已经顺利展开.一年前他的著作《分形:形,机会与维度》出版了,其中指出大量的物理、生物与数学现象都产生分形.他所研究的分形像柯克曲线一样全是自相似的.它们提出了一些有趣的数学问题,有时导致惊人的结论,以及引人入胜、高度对称的图形(蒙德尔布罗的书中展示了许多这样的图形).但在每个数学分形的例子中,都有某种内在的属性,它们在相应的现实生活的分形中并不出现.(例如英国海岸线所表现的分形行为远不如柯克曲线那样有规律.)这种高度的有序性与可预见性主要是来源于所考虑的分形在放大与平移意义下的自相似(用数学语言表述就是,它们在线性变换下保持不变).1978 年至 1979 年间,蒙德尔布罗与 IBM 的马克·拉夫(Mark Laff)开始研究在非线性变换下保持不变的分形(这种变换允许比简单的放大更为复杂的操作,包括平方、立方等等).在这种情形,要想知道相应的分形究竟是什么样子,

唯一的办法就是采用一台计算机把分维图形画出来.确实,在本世纪早期,法国的G·朱利亚(Gaston Julia)和P·法都(Pierre Fatou)曾经做过同样的工作,但不得不半途而废,部分原因就是无法画出他们所研究的对象.(蒙德尔布罗在巴黎高等工科大学读书时已知道这项工作,朱利亚曾是他在该校的老师.)

1979年底,蒙德尔布罗得出结论,认为利用计算机研究特例函数 x^2+c 的行为值得一试,这里变量 x 与常参量 c 均为复数.(我们将在后面确切说明要研究的是什么样的行为,而目前只是指出利用[85]计算机有可能绘出这种行为与参量 c 数值变化的关系图.)

具有讽刺意味的是,在关键的1978年至1980年,蒙德尔布罗却不在IBM,而是在哈佛大学访问,因此在最需要的时刻没能充分利用IBM著名的"无限的计算机设备".然而在哈佛科学中心的地下室里,他找到了一台新制造的Vax超微机,该机带有一台较旧的观察输出图象的Tektronix显示仪和一架能提供硬片的Versatec打印机.一位叫莫尔代弗(P.Moldave)的哈佛助教自愿担当(不取报酬的)程序员.这样工作便得以进行下去.

他们获得的第一张图是粗看像甲虫的双圆斑,如图20所示(图中能看到更多的细节).这正是他们所期望的——完全符合理论的预报.更令人迷惑的是一些偏离主形的小圆斑.如果对这些小圆斑作更精密的观察,就会发现它们乃是甲虫主形的小型翻版.看来我们又遇到了熟悉的分形自相似行为!通过更精密的计算获得了更精细的图形,直到图案突然呈现越来越严重的混乱.也许这是老式图象显示仪的毛病?为了证实这一点,蒙德尔布罗按同样程序在约克敦高地自己家里一台IBM计算机上进行了计算.不仅混乱没有消失,而且一张更为清晰的图象揭示出隐藏在这种混乱背后的图案.在家中作了更进一步的细致观察后,蒙德尔布罗与莫尔代弗发现有些小圆点正如他们预料的那样并不是甲虫形的翻版,而是漂亮的螺旋形图案——一族形似海马的图形(见图8与图23).蒙德尔布罗终于看到了他的新世界!

有序和混沌

古往今来，宇宙八荒，有序与混沌二者始终相互争夺霸权. 它们之间常常只有一毫之差：一点微小的压力可以使本来有条不紊的水流变成极端复杂的混沌旋涡；对安宁生活的惊扰可以使有秩序的生物(包括人口)繁衍陷入不可控制的无政府状态. 相反，混沌也可以转化为有序. 从宇宙的混沌状态进化出生物，直到出现人类，这就是一个例证. 正如我们将要看到的那样，从有序转化为混沌，以及从混沌的内部又产生出有序，这过程表现为戏剧性的形式，并可通过简单的反馈循环(feedback loops)来研究. [86]

反馈机制最本质的特征是：有一个随时间(如下例)或某个其他变量变化的量 x，x 在任一时刻的值按一定方式依赖于它前一时刻的值(见图 15). 这种类型的过程已渗透到所有的精确科学以及大多数(如果不是所有)的非精确科学之中. 对于这类过程的研究刺激了许多现代数学知识的发展，例如为了处理老 x 与新 x 之间增值为无限小量的情形而发展了许多微分方程的研究技巧.

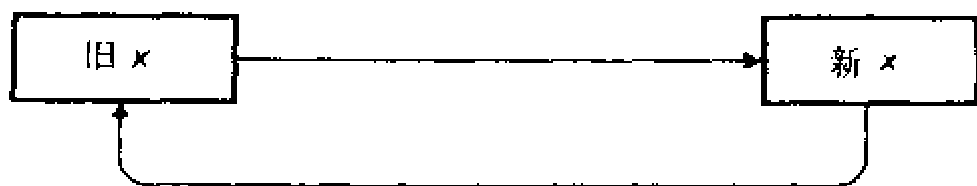


图 15 x 值变化的反馈机制.

为了从数学上研究反馈过程，我们用一个数学函数 $f(x)$ 来表示从 x 的老值产生新值的规则. 那么从 x 的一个初始值 x_0 出发，根据图 16 所示规则相继产生的 x 值为 $x_1, x_2, x_3, \cdots$. 对函数 $f(x)$ 没有必要加任何限制，虽然只有当所选择的是非线性函数时，相应的反馈过程才能引起充分的兴趣. 这里所谓的线性函数形如 $f(x) = ax + b$，a, b 为常数. 我们对 $f(x)$ 含有一个参数的情形特别关心. 参数的选择能够明显地影响相应反馈过程的行为. [87]

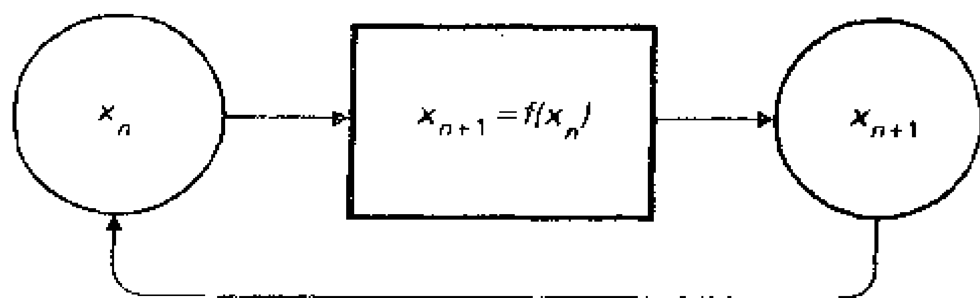


图 16 通过反馈产生 x 的相继值.

通常将反馈循环作用看作是一个动力系统,它将初始点 x_0 相继变为 $x_1, x_2, x_3, \cdots$. 由 x_0 变来的点列称为 x_0 的路径(path)或轨道(orbit). 如果路径是有序的,就称为有序动力系统,否则就称为混沌动力系统. 仅仅是这个名称就足以说明这方面的研究与大量的日常现象有关.

作为一个例子,我们来考虑若干年内人口的增长. 假定初始人口数为 x_0,设 x_n 是 n 年后的人口数,那么第$(n+1)$年的人口增长率就是

$$r=\frac{x_{n+1}-x_n}{x_n}.$$

如果年增长率为常数,则该方程对所有 n 都成立,并可改写成线性动力学定律

$$x_{n+1}=f(x_n)=(1+r)x_n.$$

n 年后的人口将为

$$x_n=(1+r)^n x_0.$$

(此表达式可以从 $x_n=(1+r)x_{n-1}, x_{n-1}=(1+r)x_{n-2}$,等等直至 $x_1=(1+r)x_0$ 逐步反推而得.)这就是所谓指数增长的例子,不仅对于人口增长,而且对于许多现实生活现象也很典型(从各方面看). 从第1章有关的论述可以明白,如果在连续若干年内不加核查,这种增长机制将会导致巨量的人口. 实际情况则是,这种增长只发生在有限的时期内,而在这段时期之后将达到一个极限. 1845 年,P·F·维哈尔斯特(Verhulst)提出了一个可解释极大人口量 X 存在的定律. 维哈尔斯特定律是说:当人口量接近 X 时,增长率将从 r 降至 0. 从数学上表述这条定律最简单的做法是将常数增长率 r 换成可变增长率 $r-$

cx_n，此处 c 为常数．因为当 $x_n=X$ 时人口增长率为零，常数 c 必须取值为 r/X．由此维哈尔斯特过程的动力学定律为

$$x_{n+1}=f(x_n)=(1+r-cx_n)x_n=(1+r)x_n-cx_n^2.$$ [88]

一旦 X 值被达到，人口就将维持常数：

$$f(X)=X.$$

若人口量比它小，就将增大；反之则会减小．如果你试算一下（手算或用计算机均可），就会发现维哈尔斯特过程描述的人口演变最终将达到一个与初始人口无关的稳定值 X．至少当 r 小于 2（即增长率 200%）时是这样——r 小于 2 的限制当然是只应用于人口增长的情况．但正如气象学家 E·N·洛伦兹（Lorenz）1963 年所指出的那样：维哈尔斯特定律当 r 值较大时描述了某些湍流现象，并且还可应用于激光物理、水力学以及化学反应理论等其他方面，因此对于 r 大于 2 时维哈尔斯特过程的行为并不是没有意义的．后来终于弄清，恰恰是在这里出现了真正迷人的结果．

设 $c=r/X$，上述规律变成

$$x_{n+1}=(1+r)x_n-(r/X)x_n^2.$$

适当改变度量单位，我们可以假设 $X=1$，于是可将规则进一步简化为

$$x_{n+1}=(1+r)x_n-rx_n^2=x_n+rx_n(1-x_n).$$

如果有计算机，你很容易进行一些计算实验，看看在最后这个方程中当 r 取不同值时维哈尔斯特过程究竟怎样变化？在每一种情况都从初始值（例如）$x_0=0.1$ 开始计算．（你的程序应对所选 r 值读出，设 $x=0.1$，将运算

$$x=x+r*x*(1-x)$$

迭代 500 次，使整个过程渐趋稳定，然后计算并打印出 x 的前 20 个或更多个值．）对小于 2 的 r 值，迭代过程很快稳定到 $x=1$ 这个平衡值．对略大于 2 的 r 值，迭代结果在两个数值之间作规则振动（$r=2.1$ 时这两个值是 0.82 和 1.13）．这种行为继续保持到 $r=2.5$，此时出现了经过四个点（0.54，1.16，0.70，1.23）的反复循环．这种情况又继续到 $r=2.55$，此时开始出现经过八个值的循环．对 $r=2.565$，循 [89]

环数值加倍,即结果在十六个数值之间反复循环,如此等等.加倍过程继续下去并越来越频繁,直到 $r=2.57$ 时,出现了无限多次加倍现象.此时动力系统的行为变得一片混乱,结果数值在整个平面上跳跃而无明显的轨迹可循.

对小于2.57的 r 值迭代收敛的各循环圈,我们称它们为吸引子(attractors).这样,$r<2$ 时吸引子由一个点组成,即定点1;$2<r<2.5$ 时吸引子是一对振动值;$2.5<r<2.55$ 时吸引子则是一个四点循环圈,等等.

[90]

画一张表明上述过程(在最初的稳定期以后)的行为与 r 值之间的关系图,即可对所发生的事情获得清楚的印象.图17的主要部分

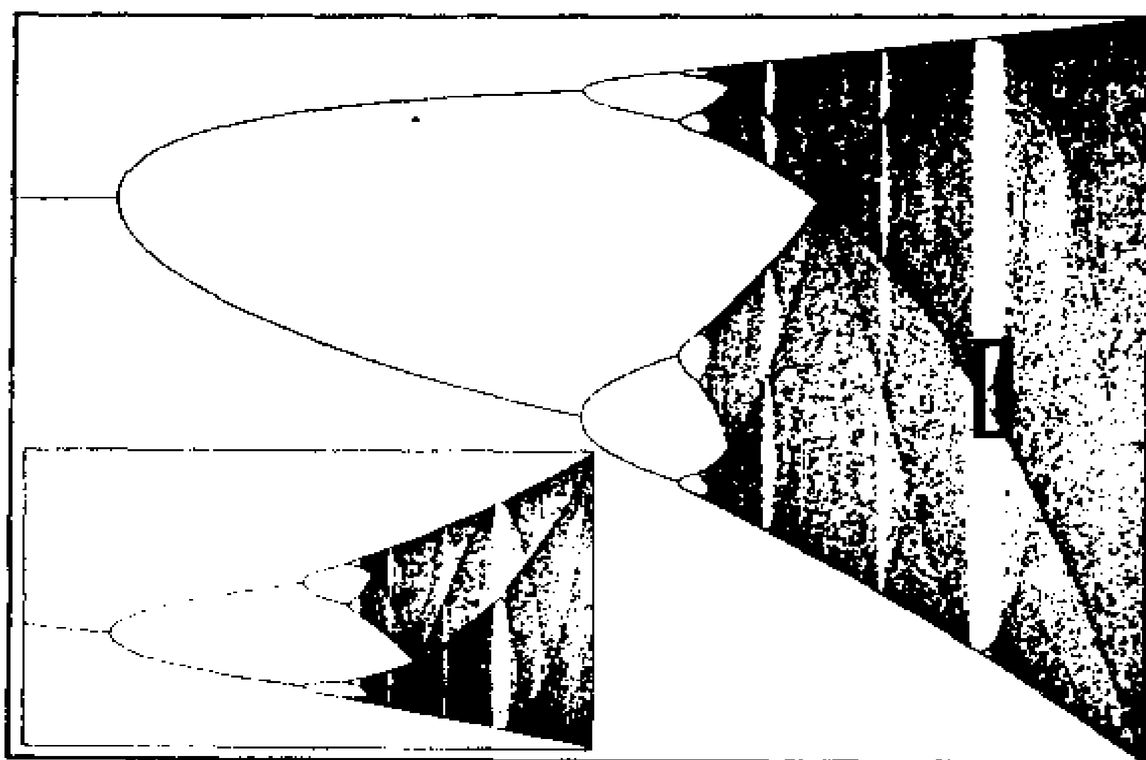


图17 维哈尔斯特过程($1.9<r<3.0$)及所示细部的放大,显示了自复制现象,r 沿水平轴从1.9至3.0取值,对每个 r 值,在垂直轴上标出经5000次迭代过程初步稳定后再连续取定的120个 x 值.对小于2的 r 值,只出现一个 x 值;对2与2.5之间的 r 值,出现两个 x 值;r 在2.5与2.55之间,出现四个 x 值;然后到2.565,则有八个 x 值.这种加倍过程越来越快,直到 $r=2.57$,此时开始出现混沌.但在这种混沌之中又开始出现有序现象,包括自复制.

表示 r 沿水平轴从 1.9 至 3.0 取值，x 经 5000 次迭代出现最初稳定期后再取连续 120 个值时的结果.

对 $r=2.57$ 以上的混沌区的仔细分析，揭示了在混沌行为的背后隐藏着许多有序现象. 例如，在 $r=3.0$ 附近，只有一个混沌带；在 $r=2.679$ 时混沌带分裂成两个；$r=2.593$ 时分裂成四个，然后是八个，十六个，依次类推，每次混沌带都加倍，一直到 $r=2.57$ 时出现无限次加倍现象，整个过程乃是动力系统本身行为的复制. 实际上存在着一个普适常数，它不仅与已经遇到的两种加倍过程而且与所有这类现象有关. 这就是所谓费根鲍姆(Feigenbaum)常数，其精确到小数后十位的近似值为

$$4.6692016609.$$

当然更神奇的是在混沌区内又出现了有序带，那里有序似乎重又占据(短暂的)统治地位. 例如在 $r=2.83$ 附近，混沌突然让位于一个三点循环吸引子. 在其中心周围的区域里，你所看到的不过是整个维哈尔斯特图象的微型复制，以混沌中央出现自身的有序带而告终. (图 17 的附图是这区域的放大，水平方向被"拉长"了.) 又是分形行为！但这还仅仅是开始！

朱 利 亚 集

前面提到的蒙德尔布罗 1980 年的工作是关于维哈尔斯特过程的研究(主要由 P·J·梅尔伯格(Myrberg)于 1960 年代进行)之继续. 蒙德尔布罗方法的主要不同之点是允许变量与常参量取复数而不仅仅是实数. 因此，代替将数变到数的过程，它将点变到点(在二维平面或阿甘得图上).

[91]

为了使问题简化，作为上述函数

$$f(x)=x+rx(1-x)=-rx^2+(1+r)x$$

的替代，蒙德尔布罗采取了较简单的公式

$$f(x)=x^2+c,$$

这也就是我们在这里要研究的函数.

假设从某个初始值 x_0(复数)开始,根据规则

$$x_{n+1}=f(x_n)$$

迭代函数 f 而产生点列 $x_0,x_1,x_2,\cdots$,让我们来看一看将会发生什么情况.维哈尔斯特过程获得的结果表明,常参数 c 的选择事关重大.首先来考察最简单的情形 $c=0$,这时动力学定律为

$$x_{n+1}=x_n^2.$$

有三种可能的输出,它们取决于 x_0 的选择.首先,若 x_0 与 O(原点)的距离小于1个单位长,迭代序列中的数将越来越小(即越来越接近 O 点),这就是说 O 是系统的一个吸引子.若 x_0 与原点的距离大于1,迭代序列中的数将越来越大,在这种情况,我们说吸引子是无限大(因为无限大并不是复平面上的一个点,"吸引子"这个词在这里完全是约定俗成的特殊用法).剩下的可能性就是 x_0 与原点的距离恰好等于1(即 x_0 位于以 O 为中心的单位圆上).在这一情形,序列的点始终不离单位圆.于是单位圆乃是两个分别由 O 和无限大控制的吸引子区域的边界.在上述例子中,复平面被一条边界曲线分成两个不同的吸引子区域,这对于蒙德尔布罗(以及其他人)研究的所有情形来说是一种典型现象.而蒙德尔布罗的发现是:对于非零参数 c 而言,不仅非无限的吸引子可由多于一个点组成,而且两个吸引子区域[92]之间的边界也可以是异常复杂和无比美丽的.

例如对于 $c=0.31+0.04i$,非无限吸引子是单独一个点,但由它和无限大所控制的区域之间的边界不是一个真正的圆,而是如图18(i)所示的漂亮的变形圆.这是一个分形变形:如果你更细致地观察边界的任一部分(以计算机为"显微镜"),就会发现熟悉的、无限重[93]复的自相似现象,而这正是分形曲线的特征!

虽然只有计算机的出现才使人们有可能深入研究这样的图形,但朱利亚和法都二人早已证明:这类图形的任何一段边界,不管它多么小,都包含了确定整个曲线所需的全部信息(就是说整个边界可以通过让这一小段边界反复经受确定系统的变换而产生,在目前的例

子中变换为 $f(x)=x^2+c$. 为了纪念朱利亚，这样的边界集现在就叫朱利亚集.

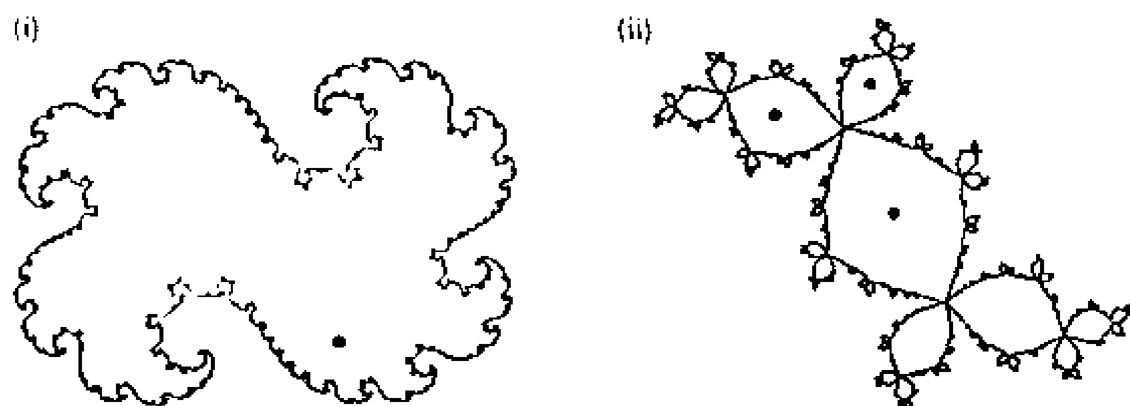


图 18 朱利亚集和它们的吸引子(详见正文).

图 18(ii)表示一个朱利亚集，与它相联系的动力学过程有一个由三点循环组成的非无限吸引子. 这里动力学定律为 $f(x)=x^2+c$, $c=-0.12+0.74i$. 图 19 则是由动力学定律 $f(x)=x^2+c$ 产生的另

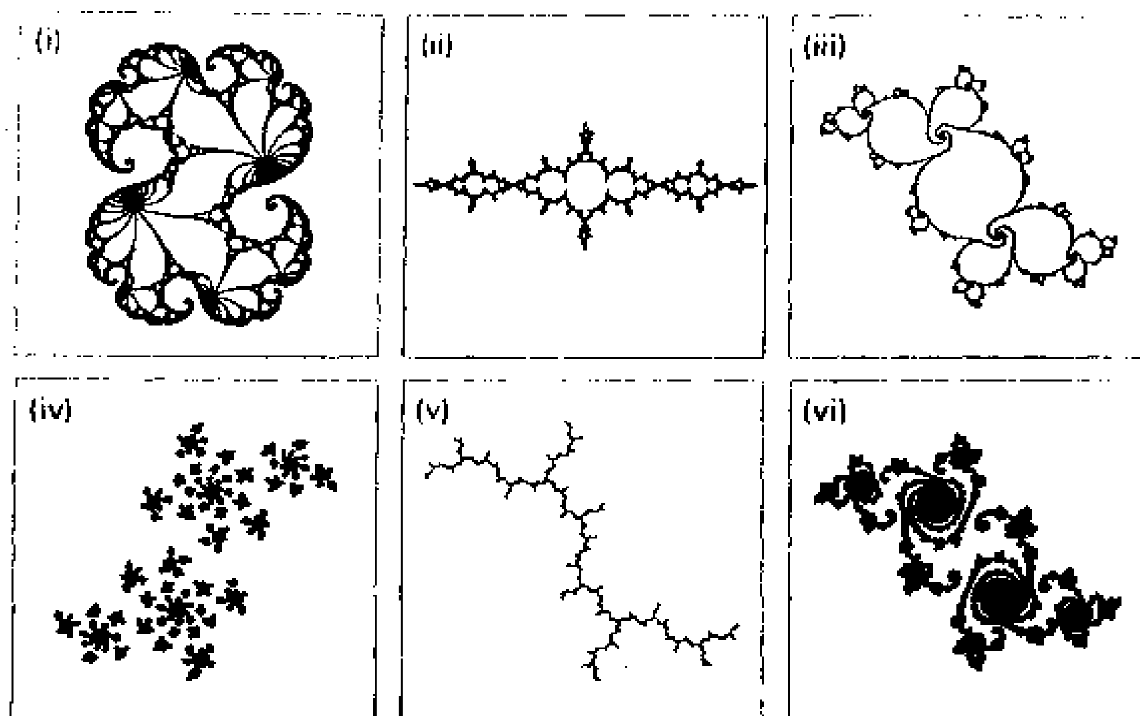


图 19 从蒙德尔布罗集边界得到的朱利亚集(详见正文).

外一些朱利亚集的例子,其中包括了区域退化为"尘点"或"树枝"的某些极端情形(详细说明见下文).

通过参数的不同选择,朱利亚集展示出丰富多采的结构,集中说明了 c 的选择是何等重要.一个自然的问题是:对于 c 的值是否存在任何识别模式,使与其相应的动力系统及其朱利亚集具有特定的形态?对这问题的研究使蒙德尔布罗在 1980 年发现了现在以他的名字命名的复平面区域(子集):蒙德尔布罗集.

蒙德尔布罗集

图 20 所示的甲虫状黑斑,就是著名的蒙德尔布罗集.蒙德尔布罗集自发现以来,已显示出与所有动力学过程行为的密切联系,不只限于此处考察的个别例子.正因为如此,蒙德尔布罗集像圆和正多边形等其他一些图形一样,在数学中占有特殊的和基本的地位.

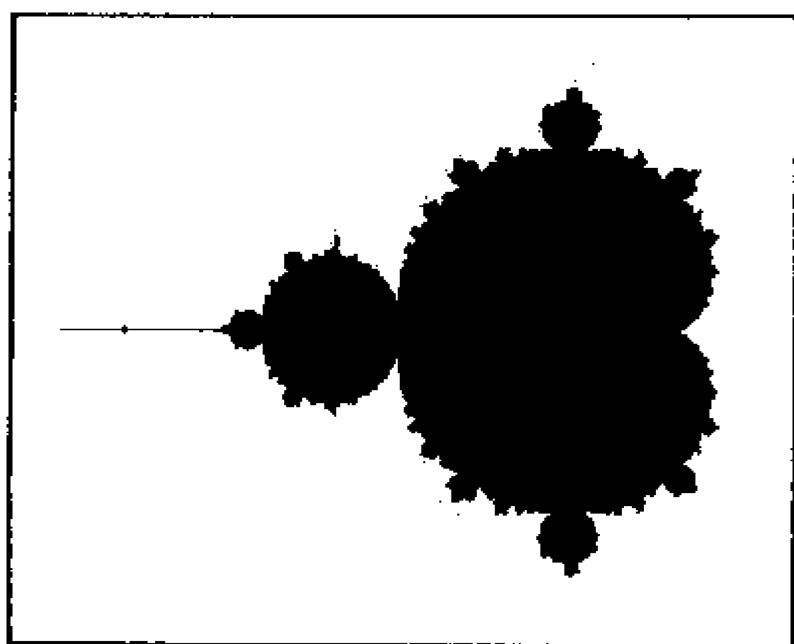


图 20 蒙德尔布罗集($-2.25 < \mathrm{Re}c < 0.75$, $-1.5 < \mathrm{Im}c < 1.5$)(详见正文).

看一下图 18 和 19 就可以清楚，一个复动力学过程或者将复平面分成一个或多个内部区域和一个伸展到无限远的外部区域（图 18(i)、(ii)，图 19（i)、(ii)、(iii))，或者产生一个退化为没有内部区域的朱利亚边界集（图 19（iv)，(v)，(vi)). 确切的行为依赖于参数 c 相对于蒙德尔布罗集的位置. 仍以过程 $f(x)=x^2+c$ 为例，我们首先考虑具有非退化朱利亚集的情形，[94] 这种情形有一个非无限吸引子.

如果 c 选自蒙德尔布罗集的主体内部，则相应的动力系统就有一个由单点(满足 $f(x)=x$ 的不动点)构成的非无限吸引子. 在这一情形朱利亚集是一个分形变形圆，如图 18(i)所示.(此时常数 c 位于蒙德尔布罗集的心脏形主体的右边缘附近.)

另一方面，如果 c 选自与蒙德尔布罗集主体相连的某个苞芽内部，此时朱利亚集由无限多个分形变形圆组成，而这些变形圆围绕着循环吸引子点或最终将变到循环吸引子的那些点. 例如图 18(ii)，c 选自蒙德尔布罗集顶部大苞芽的中心. 标出的三点形成系统的三点循环非无限吸引子. 包含该吸引子的三个区域内部的任一点将直接变向三点循环圈；而其他区域内的点则被引向最终将变到三点循环圈的所谓"局部吸引子". [95]

如果 c 是蒙德尔布罗集一个苞芽的发生点，那么朱利亚集将呈现许多卷须，这些卷须则趋向边缘稳定的吸引子，如图 19(i)所示，它们最终将稳定于一个 20 点循环圈($c=0.27334+0.00742i$)，或如图 19(ii)那样有一个四点循环圈($c=-1.25$).

最后，如果 c 是蒙德尔布罗集的其他任何边界点，其朱利亚集就是所谓西格尔盘（Siegel disc). 图 21 就是西格尔盘的一个例子（$c=-0.39054-0.58679i$). 这里存在一个被所谓不变圆（invariant circles）环绕的不动点. 在这一情形，朱利亚集所包围的区域内部的点将逐渐趋向于包含不动点的圆盘，并在其上沿不变圆围绕不动点旋转.

对于过程 $f(x)=x^2+c$ 来说，只可能有上述四种类型的朱

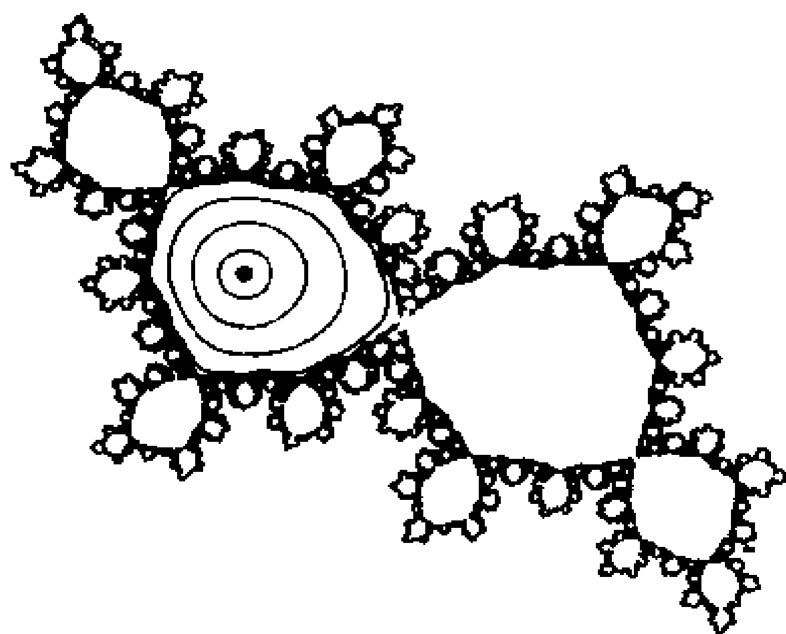


图 21 西格尔盘(详见正文).

利亚集.(1983 年，D·苏利范（Dennis Sullivan）指出：在其他复动力系统中还可能出现另一种非退化朱利亚集——赫尔曼环（Herman ring).）

关于非退化朱利亚集就介绍到这里.那么其他情形(如图 19(iv)、(v)、(vi)所示)又如何呢？蒙德尔布罗集的放大图表明它被一些发状分岔触角所包围.如果 c 正好选自这些触角之一,就可以得到类似形状的朱利亚集.图 19(v)所示为 $c = i$ 时的例子,这时系统发生的一个行为是:它只有唯一的吸引子即无限大,所有的点都将被变[96]到无限远去,除非它恰好属于发状朱利亚集本身.

图 20 并不足以看清那些触角,但从一些孤立的小点却可辨认出它们的蛛丝马迹.那么真是一些小点吗？更细致的观察(通过计算机放大)显露出它们的庐山真面目.这些"小点"原来不是别的,而是蒙德尔布罗集本身的微型复制！它们周围甚至长着更小的触角,沿着这些触角又可以发现……,如此等等,以至无穷.(在维哈尔斯特过程(图 17)中,混沌区内的空隙就相应于这类关于实轴的"苞芽"的位置.)如果 c 选自这些"苞芽"之一,相应的朱利亚集将是一个树枝集

和按相应 c 值从蒙德尔布罗集主体所得朱利亚集的无限多个复制品的组合(见图 22).

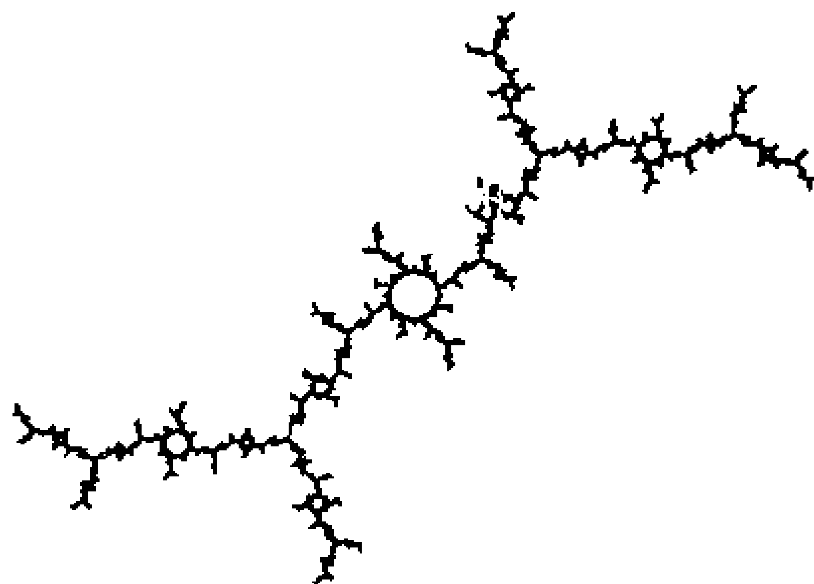


图 22 由蒙德尔布罗苞芽形产生的朱利亚集(详见正文).

现在只剩下一种可能性，即 c 选自蒙德尔布罗集(包括它所有的附连集)的外部.此时无限远是唯一的吸引子，而朱利亚集被分解成一些称为法都尘(Fatou dust)的孤立点. c 离蒙德尔布罗集越远，这些尘点就变得越来越细.若 c 选自蒙德尔布罗集边界附近一点，则尘点就变大并足以产生引人入胜的图案，如图 19(iv)、(vi)所示.(图 19(vi)与 19(iii)的 c 值接近，因此这两个朱利亚集之间也存在着令人注目的相似.)这种尘点图案与混沌动力学一样总是呈现分形(即自相似)的性质.

毫不奇怪，由于蒙德尔布罗集的边界在相关的系统动力学中扮演如此重要的角色，这边界本身就引起了巨大的兴趣.正如人们所期望的那样，现已弄清：这个边界区域本身具有复杂的分形外表.图 23 [97] 是对这个不可思议的世界的一瞥.这是一个寸土必争的世界，是一个只有通过计算机才能认识的世界——认识的程度完全依赖于计算机

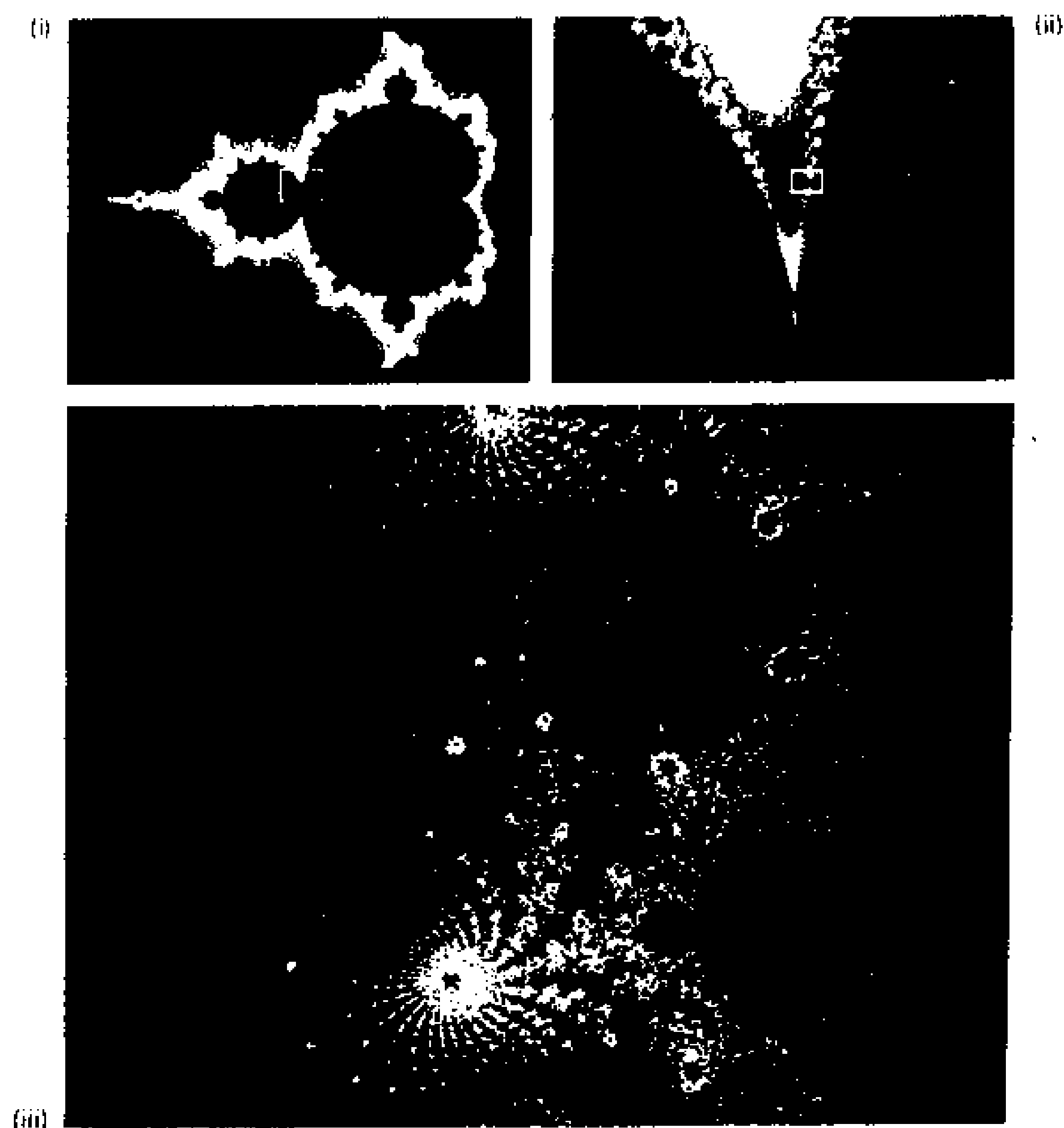


图 23 蒙德尔布罗集边界区域巡游.

的能力.如果说有什么真正属于计算机时代的数学分枝,这无疑就是[98]一个.

阅读文献

要想真正欣赏本文描述的世界,必须借助彩色图象,只有色彩才能提供反

映系统动力学的鲜明的轮廓图.一本包含彩色与黑白图片(以及说明)的杰出著作是 H.O.Peitgen 和 P.H.Richter 合编的 The Beauty of Fractals(Springer - Verlag,1986).该书非常值得推荐.

关于分形几何及其应用的最新介绍,可参阅 B.Mandelbrot 的 The Fractal Geometry of Nature(W.H.Freeman,1982).

(李文林译)

第5章 单 群

庞大的定理

1980年夏，俄亥俄州立大学的数学家R·所罗门（Ronald Solomon）在解决了代数中的一个技巧性很强的问题后，放下了笔. 随着这个简单的动作，一个在本世纪40年代提出的疑问终于烟消云散；要知道，这个问题曾使一百多位来自美国、英国、德国、澳大利亚、加拿大和日本的数学家为之奋斗. 所罗门的成果恰好为解决这个庞大而又高度复杂的难题添上了最后的一笔. 这个问题就是有限单群的分类①.

分类定理是迄今为止最庞大的数学定理. 最初的证明几乎长达15000页，分散在约500篇刊于数学期刊的文章中，有一百多位数学家为此作出了贡献. 随着研究的进展，许多新发现导致了计算机算法理论、数理逻辑、几何和数论等学科的进步. 有人一直在考虑它可能会对物理中统一场论的构作有所帮助.

我们的讨论将涉及数学中很深刻的结果，但故事的开端却十分平凡：我们熟知公式

[100]

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

是二次方程

① 所有专门术语将在适当的时候给予解释. ——原注.

$$ax^2+bx+c=0$$

的根的表达式.人们试图对次数更高的方程(即方程中 x 的幂次大于 2,诸如下面的三次方程)得到类似的解.(此处"类似的解"是指公式中包含加、减、乘、除以及开方根这些基本的代数运算.这种解有时称为根式解.)

É·伽罗华

考查古代泥板文书可清楚地看出,生活于公元前 1600 年的巴比伦数学家已知道如何解二次方程,尽管他们没有使用我们现在的代数符号去表达方程及其解.形如

$$ax^3+bx^2+cx+d=0$$

的三次方程的(根式)解,直至 16 世纪才被发现,那时意大利数学家 S·de·费罗(Scipio de Ferro)和 N·丰塔那(Nicola Fontana)彼此独立地找到了解法.1545 年,G·卡尔达诺(Girolamo Cardano)在他的《大术》(Ars Magna)一书中公开发表了丰塔那的方法.这部书还讲述了 L·费拉里(Ludovico Ferrari)求解四次方程的方法(将其约简为三次方程).但事情的发展似乎就此了结了.虽然有很多数学家作出了努力,其中包括 18 世纪中叶的伟大的瑞士数学家 L·欧拉(Leonhard Euler),但没有一个人能找出五次方程

$$ax^5+bx^4+cx^3+dx^2+ex+f=0$$

的解的公式.J·L·拉格朗日(Joseph Louis Lagrange)在 1770 年时提出:大概不可能有这种解(即指根式解);1824 年,挪威数学家 N·H·阿贝尔(Niels Henrick Abel)证明了拉格朗日的看法.

[101]

如果不存在一般的方法(即没有解的公式)去解五次方程,那么自然会问:是否有什么方法能够来确定一个给定的五次方程能否有(根式)解?阿贝尔去世(1829 年,年仅 26 岁)前正竭尽全力地研究这个问题.在这一时期,碰巧还有一位年轻人也在勤奋地钻研这个问题,而且最终取得了成功,他就是埃瓦里斯特·伽罗华(Évariste Ga-

lois).可是这位年轻人获得的非凡成果,在他因决斗去世后11年才开始得到数学界的承认;其间风云变幻,人事繁杂,而能够从中带着巨大的荣耀脱颖而出的只有年轻的伽罗华一人.

伽罗华1811年10月降生于巴黎近郊.14岁那年因考试不及格而重上三年级,这倒成了他对数学发生兴趣的契机.他发现数学能帮助他逃避其他课程给他带来的无聊.不幸,他对数学不断增长的热情,招致他的课上课下作业的成绩越来越糟;当他15岁参加声望很高的巴黎高等工科大学的入学考试时,伽罗华失败了,不得不进入较普通的师范学校.就是在这所学校,伽罗华于次年写出了他的第一篇数学论文——虽不十分出色但已显示了他的能力,内容涉及连分数.良好的开端预示着希望,可惜他很快就陷入了一系列不幸的事件,最终不得不完全放弃他所如此热爱的学科.

他的下两篇论文(研究多项式方程)遭到法国科学院的拒绝.更糟的是,两篇论文手稿还莫明其妙地被丢失了.1829年7月,他在巴黎高等工科大学的入学考试中再次失败;这次失败很可能跟他在回答考官提出的一个特殊问题时的表现有关.当问到他"算术对数"理论的梗概时,伽罗华(完全准确但极不得体和通融地)答道并不存在算术对数.怀着沮丧之情,伽罗华于1830年初又向科学院提交了另一篇论文,这次是为竞争一项数学大奖.科学院秘书傅立叶(Fourier)将其手稿拿回家去审读,不料在写出评审报告前去世了,此文再也没有找到.三失手稿,加之考巴黎高等工科大学两度失败,伽罗华遂对科学界产生排斥情绪,变成了今日所谓的学生激进分子.他被学校开除,被迫去寻找私人辅导教师的工作以谋生路.虽然他当教师并不十分成功,他的数学研究工作却仍相当活跃;他在这一时期写出了将成为他最著名的论文"关于方程可根式求解的条件",并于1831年1月送交科学院.[102]

递交这篇论文是他最后一次尝试让数学界承认他的工作. 到3月, 科学院方面仍杳无音讯, 于是他写信给院长打听他的文章的下落, 结果又如石沉大海. 最后他放弃了一切希望. 他可能再没有去

做数学；改而参加了国民卫队（一种共和主义者的组织）. 在这里，他跟在数学界时一样运气不佳. 他刚加入不久，卫队即遭控告阴谋造反而被解散. 在5月9日举行的一次抗议聚宴上，伽罗华手中举着出鞘的刀提议为国王干杯，这一手势很自然地被同伙们解释成是要国王的命；第2天他就被捕了. 在审问中，他自称他实际上说的是"为路易·菲利普干杯，如果他成了卖国贼"，可后半句话被当时的骚乱声浪淹没了. 不论此话是真是假，反正他被判无罪，并于6月15日获释.

7月4日，他终于打听到他给科学院的那篇论文的命运：因"无法理解"而遭拒绝，审稿人泊松(Poisson)是这样结束其评审报告的：

> "我已尽了一切努力去理解伽罗华的证明. 他的推理不够清晰，不够充分，我们无法判断其正确性；本报告也不能就此提出任何想法. 作者宣称，该文研究的特殊对象是具有众多应用的一种更普遍的理论的一个组成部分. 也许，整个理论的各个不同的部分能相互澄清，因而比孤立的部分更容易掌握. 我们不妨建议作者发表其完整的结果，以便得出明确的意见. 但就目前他送交科学院的部分结果而言，我们不能推荐说应给予承认."

站在恩赐立场上提否决意见的鉴定人，可能感到他的这个报告无可挑剔，我们也不知道它对伽罗华其后的行为有无影响. 但是在7月14日他又遭逮捕，因为他在公共场所身着已被解散的国民卫队的制服. 这次他被判了六个月的监禁.

在获假释不久，他陷入了与某位斯特凡妮小姐的恋情.（我们不知她姓什么. 在伽罗华的一份手迹中出现过她的姓，但被重重地涂抹掉了，可能是她拒绝伽罗华的证据.）这导致了他的早亡. 这次恋爱事件不知何故引出了一场决斗（小仲马(Alexandre Dumas)认为这场决斗其实是政治原因促成的变相行刺阴谋）. 5月29日，决斗的前夜，伽罗华写了封长信给他的朋友A·舍瓦利耶(Auguste Chevalier)，其 [103]

中大致描述了他的数学理论，从而给数学界留下了唯一一份它将蒙受何等损失的提要.在第二天的决斗中（离 25 步远用手枪射击），伽罗华的胃部中弹，24 小时后去世.

他遭拒绝的是篇什么样的论文？1843 年 7 月 4 日，J·刘维尔（Joseph Liouville）在法国科学院演说的开场白这样说：

> “我希望我的宣告能引起科学院的兴趣；在埃瓦里斯特·伽罗华的那些文章中，我已发现如下漂亮的问题的一个既精确又深刻的解答：……是否根式可解？……”

伽罗华留给世界的最核心的概念是群，这对所有时代都是最有意义的概念之一，在许多数学领域有它的应用，而且可用于物理、化学和工程学分支.

这是个完全抽象的概念.它之所以有如此威力，原因是有大量群的实例存在，它们往往各具不同的特性.群的概念具有多面性，所以可用各种方式介绍它.此处选择的方式将利用到平面几何图形的对称性，这纯粹是因为它提供了易理解的形象直观的例子.我们在本章较后的部分将遇到其他类型的群，它们跟我们即将讨论的对称群一样正统.

对　　称

考虑如图 24 所示的等腰三角形.用普通的日常用语讲，这种几何图形对图中的垂线是“对称的”.我们说三角形 ABC 是对称的，意指三角形在垂线左边的部分（即较小的三角形 ABD），是右边部分（即 ACD）的镜像，想象中的镜子沿垂线 AD 摆放，并垂直于三角形所在的平面.如果将此图形的两半调换（或者说反射）一下，整个图形跟原来的丝毫不差地出现在相同的位置上，只是 AB 和 AC 互换，BC [104] 翻转了方向.

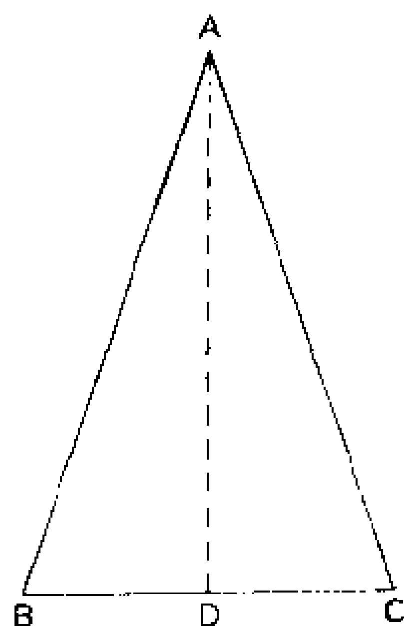


图 24　等腰三角形的对称.

一般而论，对于平面上的任意一个图形 S 和平面上任意一条直线 l，S 关于轴 l 的反射是指一种行为(一种作用)，它将 S 上的每个点移至它对 l 的镜像处，所谓镜像则是过该点并垂直于 l 的直线上的某个点，后者到 l 的距离跟该点到 l 的距离相等. 注意，对图形进行变换的行为是指反射这种作用，而不是指行为的结果.(我们将集中讨论行为而不是其结果，理由将在下面澄清.)对一个图形 S 所作的反射得出的图形被称为 S 在该反射下的像. 图 25 显示了某些反射(的结果)的例子.

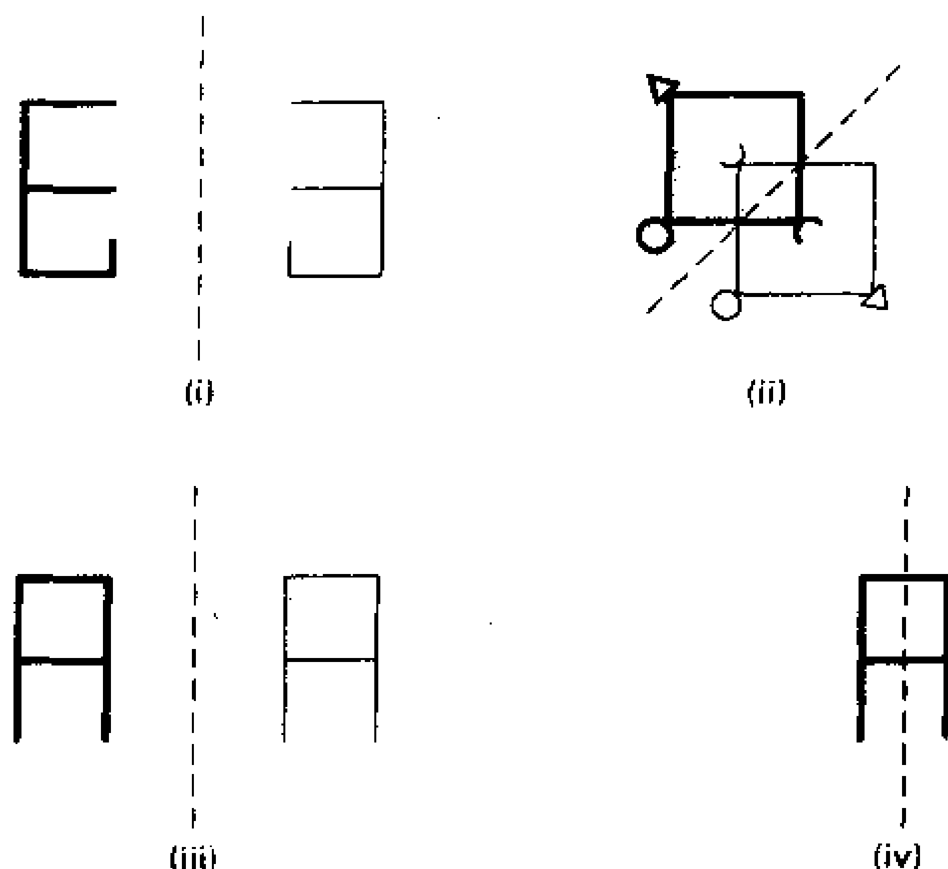
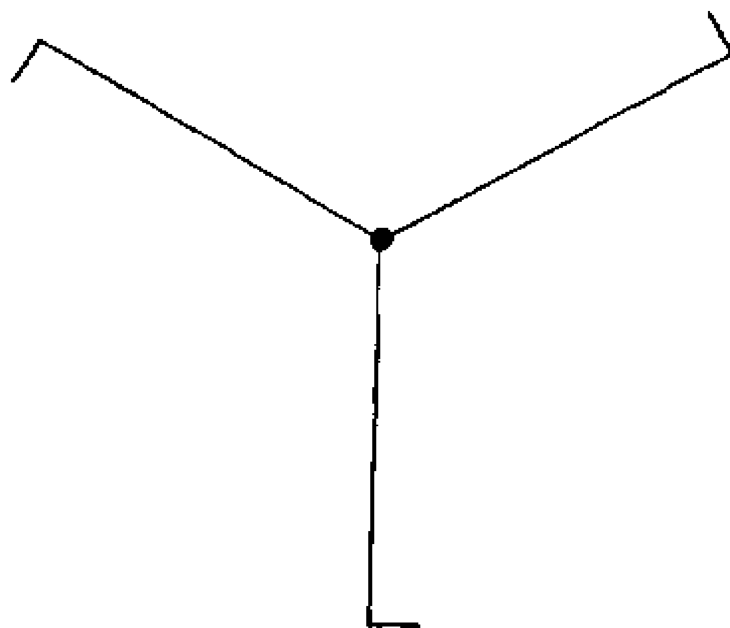


图 25　关于轴的反射. 图中以虚线表示轴，用粗线表示被反射的图形，反射后得到由细线表示的像. 在(iv)中，像跟原图形重合.

[105]

利用反射这一概念，数学家称平面图形 S 是关于直线 l 对称的，若 S 关于 l 的反射得到的像跟 S 在该平面上占据完全相同的位置．此时 l 称为对称轴．（图形 25 中的（iv）是关于轴对称的例子．）轴对称有时也称为*左右对称*．

左右对称就是日常生活中（对平面图形）使用“对称”这个词的含义，但数学家还有一类对称，它可用图 26 来说明．图中所示的形状绕中心点旋转 120°角后，它在平面上占据的位置完全没变．这是旋转对称的例子．（注意，这里我们考虑的是关于*点*的旋转．若将图 24 中的三角形作关于直线 AD（把它当作轴）的 180°的旋转，所得图形仍占据原来同样的位置，而且跟关于 AD 的反射的效果一致．）



[106]

图 26 旋转对称

群

再次考虑图 24 中的等腰三角形．它有多少种对称呢？我们问的是，都有哪些（关于轴的）反射和（关于点的）旋转，它们作用于三角形

所得的像跟原三角形占据完全相同的位置？显然，关于垂线 AD 的反射是一个．让我们用字母 r 标记这个反射．还有其他的对称吗？很明显再没有其他的反射对称了，但旋转对称呢？关于任一个点作360°的旋转将使图形回到起始状态，但这不能算，因为此时所得的结果根本没出现任何变化（相反，在作反射 r 时发生了真正的变化，点 B 和 C 最后占据了跟开始时不同的位置）．因此，我们本可不考虑这种平凡的情形——或至少要另眼对待它．但正如考虑数0（加零并不改变原数）和数1（它对乘法不起作用）是有用的一样，将恒同变换 I 看作是一种对称也是必要的，它使平面上的任何点都不改变位置．I 可看作是0°旋转．

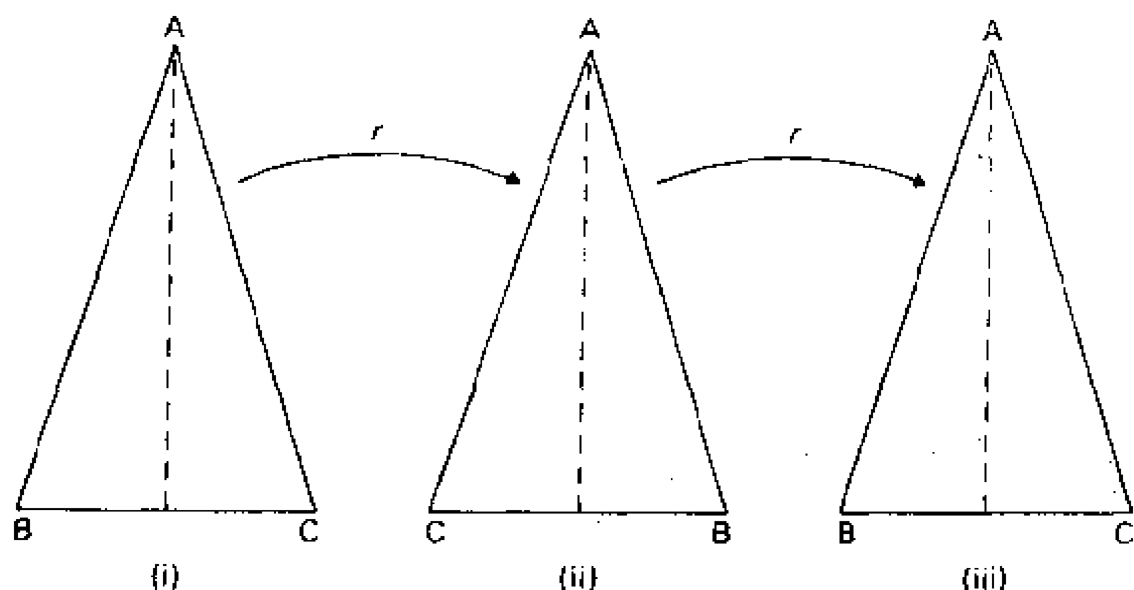


图 27　相继进行的反射．

[107]

假定我们对三角形 ABC（见图27(i)）施行反射变换 r，则导出如图27(ii)所示的三角形 ACB．如果再对 ACB 施行 r 将出现什么结果呢？显然我们将再次回到最初的状态——图27(iii)．所以，相继实行两次 r 的后果跟不作任何变换相同（对后者，我们可以表示为施行一次恒同变换 I）．我们可以用符号表示为

$$r * r = I,$$

其中的星号 * 意为"再次施行"．（因此，若 a 和 b 为两个对称，$a * b$

意为先施行 a，再对结果施行 b 的运算.）利用同样的记号，我们可以描述施行其他一些对称的作用：

$$r * I = r,$$
$$I * r = r,$$
$$I * I = I.$$

这四个等式可用下表加以总结：

等腰三角形：

*	I	r
I	I	r
r	r	I

为了观察施行一个对称 x 后接着施行另一个对称 y 的效果，你先顺着表中第 x 行看，直到第 y 列时止，此时相应的项是 $x * y$，即两个对称组合后的结果.

注意，上述讨论暗含了相继施行两个对称本身也是一个对称. 情况确实如此（你想一想就会明白这是显然的）.

当我们对图 26 所示的三叉状物进行同样的分析又将如何呢？[108] 此时有三个对称：作 120°的逆时针旋转（称为 v），作 240°的逆时针旋转（称为 w），以及恒同变换 I，此时一切保持不变.（你可能会问"顺时针时情况如何？"回答是作 120°的顺时针旋转跟 w 相同，作 240°的顺时针旋转等同于 v，因此我们已考虑了所有可能的情形.）因为相继作两个 120°的旋转效果跟作一个 240°的旋转相同，因此显然有

$$v * v = w.$$

同样，两个 240°的旋转等同于一个 120°的旋转，因此

$$w * w = v.$$

全部相继的对称由下表给出：

三角架：

*	I	v	w
I	I	v	w
v	v	w	I
w	w	I	v

再举一例. 等边三角形（见图 28）有六个对称. 它们是恒同变换

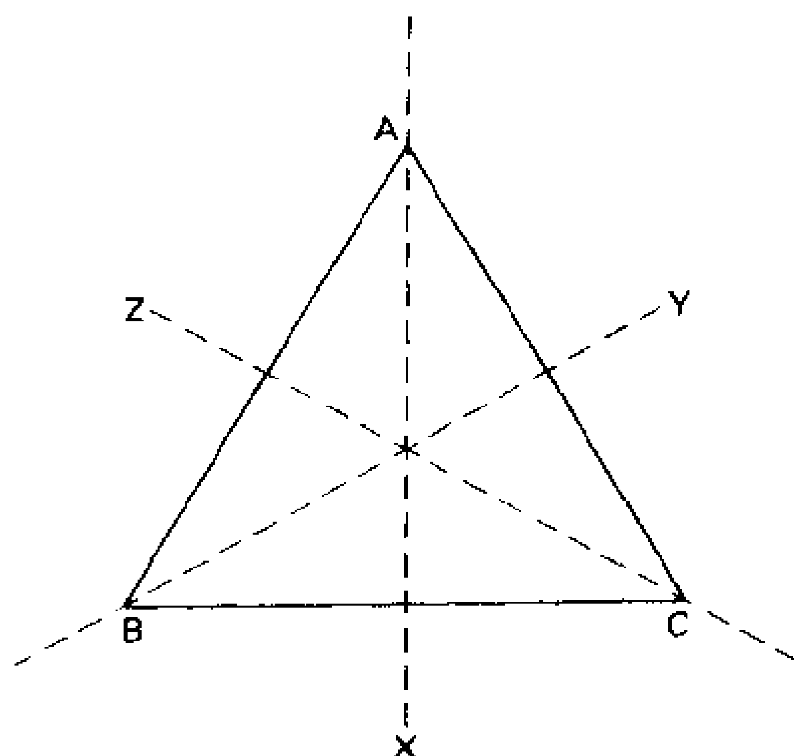


图 28 等边三角形的对称.

I,分别作 120°和 240°的逆时针旋转 v 和 w,分别作关于直线 X,Y,Z 的反射 x,y,z.(三角形变动时直线 X,Y,Z 保持不动.)这些对称的组合由下表给出：

等边三角形：

*	I	v	w	x	y	z
I	I	v	w	x	y	z
v	v	w	I	z	x	y
w	w	I	v	y	z	x
x	x	y	z	I	v	w
y	y	z	x	w	I	v
z	z	x	y	v	w	I

[109]

如果你想验证表中的各项,不妨用硬纸片剪一个等边三角形,三个角分别标明 A,B,C,然后把它放在画有三条直线 X,Y,Z 的一张纸上.这样你就可以实地施行各种旋转和反射.(在进行反射时三角形的两面都要标上字母.)

两个以上的对称进行组合时的情形又如何呢？这不必特别加以

考虑,因为不管应用多少个对称,都等价于相继进行两个对称的组合.例如,对于刚考虑过的等边三角形,$(x*y)*v$ 跟 $v*v$ 是一样的,它又正好是 w,这里我们两次使用了那张表.为了指明对称组合的次序,上面的括号是必须的:y 在 x 之后施行,然后再施行 v.另一种组合 $x*(y*v)$表示:先施行 x,然后再施行 $y*v$.如果你按第二种方式组合,将得到什么结果呢?你知道,$y*v$ 等于 z,所以 $x*(y*v)$等价于 $x*z$,后者又等于 w,这跟按第一种方式组合得到的结果相同.仔细思考一番,你将发现这并不奇怪.对称的所有组合都具有这种性质:若 a,b,c 是有关某个图形的对称,那么

$$(a*b)*c=a*(b*c).$$

[110] (这一性质称为运算 * 的结合律.)

在给"群"下定义之前,我们还需要进行最后一项考察.如果你有一个对称,并"逆向施行"之,显然它构成另一个对称.(就反射而论,"正向"和"逆向"施行并无差别.对于旋转,只要改变旋转方向即可.)一个对称 x 的"逆向施行"称为 x 的逆,通常记作 x^{-1}(读作"x 的逆").对于等腰三角形,你有 $r^{-1}=r$(这对任意的反射都成立).对于三角架,你有 $v^{-1}=w$ 以及 $w^{-1}=v$;对于等边三角形,则有 $v^{-1}=w$,$w^{-1}=v$,$x^{-1}=x$,$y^{-1}=y$ 和 $z^{-1}=z$.$I^{-1}=I$ 则对所有情形都适用.你从以上的分析中发现了些什么吗?如果你就每一种情形去对照一下那些表的话,你将注意到 $a^{-1}*a=a*a^{-1}=I$ 永远成立.这也不是偶然的,你只要想一想何为恒同对称,何为一个对称的逆,就会明白其中的缘由.

现在你应该能隐约地感到上面的讨论中存在某种熟悉的东西了,即使你过去从未考虑过对称.如果不考虑零(它没有逆),它们不是跟普通的有理数(分数)乘法十分相像吗?任意两个非零的有理数的乘积是另一个有理数;乘法允许任意编组进行(即对任意的数 a,b,c,有$(ab)c=a(bc)$);所有非零的有理数 x 都有跟它相逆的有理数 $x^{-1}(=1/x)$,使得 $xx^{-1}=x^{-1}x=1$,此处的 1 具有乘它不改变乘积的性质.也许你没有思考这些问题,你脑中的例子是整数的加法:

两个整数的和是另一个整数；$(a+b)+c=a+(b+c)$永远成立；存在恒等数0，加上它不改变原来的数；每一个整数 x 都有它的逆 $(-x)$，使得 $x+(-x)=(-x)+x=0$. 也许你心中还有其他例子. 事实上例子多得很，包括伽罗华在研究解五次方程问题时所思考的实例. 而所有这些例子都属于伽罗华提出的一般的"群"的概念的特殊情形.

对数学家而言，群由下列要素组成：

(1) 一个集合 G 以及

(2) 一种运算 $*$，它为 G 中的任意一对元素 x 和 y 确定一个仍属于 G 的元素 $x*y$.

[111]

运算 $*$ 需满足下列三个条件("群公理")：

(3) 它满足结合律：对 G 中任意的 x,y,z，有

$$(x*y)*z=x*(y*z).$$

(4) G 中存在单位元 I，使得对 G 中任意的 x，有

$$I*x=x*I=x.$$

(5) G 中的每个成员都有一个逆：若 x 属于 G，则存在 G 中的元素 y，使得

$$x*y=y*x=I.$$

我们已见到过几个群的例子. 若 G 是平面上某给定图形的所有的对称形成的集合，$*$ 是相继施行两个对称的运算，这就是一个群. 又若 G 是非零有理数的集合，$*$ 是普通的乘法，它也是一个群. 再若 G 是整数集合，$*$ 是普通的加法，它又是一个群. 这两个由数构成的群中的运算都满足交换律，即对群中任意元素 x 和 y，有

$$x*y=y*x.$$

但这并不是要求群必需满足的性质. 例如，等边三角形的群就不满足. 你不妨去看一下此群的表，你将发现 $x*v=y$ 而 $v*x=z$. 在其中 $*$ 是可交换的那些群，通常称为阿贝尔群(以前面提到过的挪威数学家命名).

尽管群中的运算 $*$ 没有必要是人们已熟悉的运算，如算术运算，

但一经采用通常就用原来的名字称呼它(乘法,加法,或其他可能的名字).不过,如群中定义的运算原来并不知道,或没有通用的名称,我们一般就称这样的 $*$ 为群的乘法,称 $a*b$ 为 a 和 b 在群中的积.
[112] 这样做纯粹是为了方便,不要从这种称呼本身望文生义而妄加推断.

因为不同的群的例子非常多,所以群的概念十分有用——它不仅在数学中几乎无处不在,同样也出现在其他科学中.晶体中的周期性,原子中的对称,基本粒子的交互作用,它们都涉及到群.凡对一般的群成立的论断,对任一特殊的群也成立.那么,数学家是如何建立任意的、抽象的群的性质的呢?回答是从群的定义出发,任何命题都必须用严格的数学方法加以证明.

作为例子,我们将证明群中的任一元素必定恰好有一个逆.(这是在使用记号 x^{-1} 表示 x 的逆之前必须证明的.)群的定义中的条件5保证群中的元素至少有一个逆,但并不排除有多个逆.当然,在前面给出的群的例子中,显然没有一个元素有多余一个的逆,但这对我们需要的证明无所裨益,我们的证明应适用于所有的情形,包括我们从未考虑过的可能存在的群的例子.

下面就是证明.设 G 是任意一个群,设 x 是 G 中的任一成员.又设 y 和 z 是 x 的两个逆.我们的目的是证明 $y=z$.因 y 和 z 都是 x 的逆,据条件5应满足:

$$x*y=y*x=I, \tag{3}$$

$$x*z=z*x=I. \tag{4}$$

据条件4,我们有

$$y=I*y.$$

利用等式(4),可得

$$y=(z*x)*y.$$

由条件(3)即得

$$y=z*(x*y).$$

再利用等式(3)得到

[113]

$$y=z*I.$$

于是，对 z 使用条件 4 即可推出

$$y = z.$$

证明完毕. 请注意由定义赋予群的所有结构性条件，在上述证明中是必不可少的.

你能想象群论中的大多数证明，跟上面极其容易的例子相比要复杂得多（因而通常并不如此详尽地写出证明步骤），它们常常涉及其他概念；但是所有的证明有一个共同的特征，即整个证明过程全部由基于最初的假设的逻辑推理步骤组成.

群的其他例子

上面考虑的由对称组成的群都是关于平面图形的. 但同样的想法也适用于三维空间中的立体图形. 例如，立方体有 24 个旋转对称（这里是关于一个轴的旋转，而二维情形是关于一个点的）. 为了说明这点，我们注意立方体的每一个顶点可以移动到另外任一顶点处，由顶点引出的三条边可以用三种方式转动，如果再把反射包括在内（是相对于一个平面，而不是相对于一条线的反射），则立方体共有 48 个对称.

十二面体是由十二个全等的正五边形组成的类似球形的立体（见图 29），它有 60 个旋转对称（如包括反射则有 120 个对称）. 立方体和十二面体两者的旋转对称自身都构成群，它包含于图形的全部对称之内. 这时，数学家们称旋转对称构成了全部对称群的子群.

十二面体的旋转对称构成了最小的非交换单群（见后文），伽罗华利用了这种群的单纯性及其元素个数不是素数这一事实，证明了一般的五次多项式方程不可能有根式解.

到目前为止，我们所见的例子中的群有些是无限的（如整数加法），有些是有限的（如各种对称群），我们在本章将主要关注有限群，矩阵则提供了有限群和无限群的例子. 114

一个矩阵是一长方形数组（在我们目前的例子中，其中的数可以

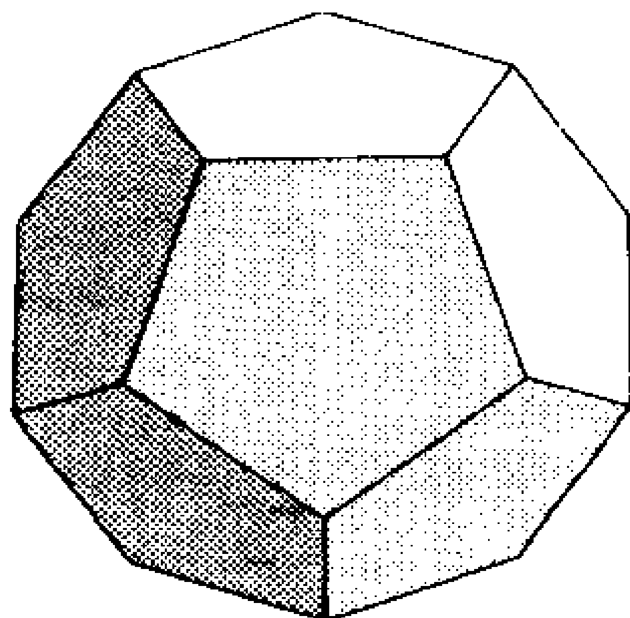


图29 十二面体

是有理数或实数).通常将它写在一个括弧内.例如:

$$\begin{pmatrix} 1 & 3 & 3/4 \\ 2 & 1/4 & -5 \end{pmatrix}.$$

矩阵的形状和大小可以是任意的,但我们将特别关心方阵,它的行数与列数相同(这个数通常称为矩阵的阶).下面是一个二阶(方)矩阵的例子:

$$\begin{pmatrix} 21 & -5 \\ 3.8 & 20 \end{pmatrix}.$$

矩阵有其自身的算术运算,两(同阶)矩阵相加是很方便的,只须将对应的项相加,如:

[115]

$$\begin{pmatrix} 1 & 3 \\ -2 & 6 \end{pmatrix} + \begin{pmatrix} 2 & 5 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 8 \\ 1 & 7 \end{pmatrix}.$$

乘法有一点儿复杂.简而言之,就是将第一个矩阵的各行与第二个矩阵的各列项对项地乘,再将它们加起来.对于二阶矩阵,可用一个代数的例子和一个数字的例子给以最清晰的解释:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \times \begin{pmatrix} v & w \\ x & y \end{pmatrix} = \begin{pmatrix} (av + bx) & (aw + by) \\ (cv + dx) & (cw + dy) \end{pmatrix},$$

$$\begin{pmatrix} 1 & 3 \\ -2 & 5 \end{pmatrix} \times \begin{pmatrix} 2 & 4 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} (2+9) & (4+3) \\ (-4+15) & (-8+5) \end{pmatrix}$$
$$= \begin{pmatrix} 11 & 7 \\ 11 & -3 \end{pmatrix}.$$

第二个例子说明矩阵乘法是非交换的(显然加法是交换的):

$$\begin{pmatrix} 2 & 4 \\ 3 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 3 \\ -2 & 5 \end{pmatrix} = \begin{pmatrix} (2-8) & (6+20) \\ (3-2) & (9+5) \end{pmatrix}$$
$$= \begin{pmatrix} -6 & 26 \\ 1 & 14 \end{pmatrix}.$$

你可能要问:为什么要这样复杂地定义乘法呢?为什么不像加法那样简单地以对应项相乘呢?这是数学家们研究了矩阵的一些应用(特别是解很大的联立方程组)后,根据这些应用的需要而给出的定义.今天,矩阵演算是如此重要,以致于任何计算系统,无论是用于科学还是商业的都理所当然地配备了处理矩阵演算的软件.确实,矩阵演算大概是今日计算机最常进行的一项数值工作.

三阶或更高阶的矩阵的加法及乘法的定义可用类似于上面所给的二阶定义给出.实际上,我们以后将要说到的每件事都可应用到任意大的阶的情形(仅须作一些显而易见的修正),但为了清楚起见,我们将继续集中讨论阶为 2 的矩阵.

[116]

阶为 2 的矩阵(或任意阶的矩阵)按加法构成一个群.两个矩阵之和仍是一个矩阵,加法满足结合律,存在一个单位矩阵

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

且任一矩阵的逆可通过在每个元素前面加一负号而得,这个群还是可交换的.

乘法的情形如何——是否也给出一个群?确实,两个矩阵的乘积仍是一个矩阵,矩阵乘法也满足结合律(这不是非常显然的,但如动手作一下推导,你就会看出这是对的).存在一个单位元,即矩阵它

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

乘以任何一个矩阵都使后者保持不变.至于是否存在逆的问题,通过

直接计算，可以验证

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \times \begin{pmatrix} d/H & -b/H \\ -c/H & a/H \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

其中 $H = ad - bc$. 两个矩阵左右交换相乘也是如此，故而矩阵

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

的逆矩阵应是

$$\begin{pmatrix} d/H & -b/H \\ -c/H & a/H \end{pmatrix}.$$

但前提应是第二个矩阵存在！当 H 是 0 时就不行了(0 不能作除数). H 不为 0 的矩阵是有逆的，这样的矩阵叫做可逆的(或有时称为非奇异的). H 为 0 的矩阵没有逆，并称为奇异的(或不可逆的). [117]

由于存在奇异矩阵，它没有逆，矩阵在乘法下不构成群. 但如果我们只考虑可逆矩阵，就得到一个群. 这显然吗？未必. 结合律是没有问题的，因为它对所有矩阵都成立，无论它是否可逆. 由于单位矩阵是可逆的(因此是我们考虑的集合中的一个成员)，所以我们有了群中的单位元. 元素的逆呢？我们选择的集合中的元素都是可逆的，但你必须检查这些逆本身也在这个集合中. 这是很容易的事，但必须核实，最后还要检验一件事，即你必须知道，两个可逆元素相乘的积仍是可逆的(即仍在我们的集合中). 这也是可用代数演算证明的简单的事(你自己试试看).

可逆矩阵构成的群显然是无限的(我们早已指出它是非交换的). 后面我们将考虑的非常重要的群是由可逆矩阵群的有限子群构成的.

另一类重要的群叫做钟群. 最显然的例子由 12 小时的钟点给出. 取从 1 到 12 的整数集，利用“钟表加法”作演算，其中的 12 在计算中当作 0，过了 12 你又从 1 数起，例如 5 加 5 是 10，7 加 8 是 3，11 加 11 是 10，7 加 12 是 7，等等. 这样给出一个群，其单位元是 12. 群中任一数的逆是 12 减去该数所得的差，所以 7 是 5 的逆，9 是 3 的

逆，等等.

这里12这个数没什么特别的，任一个数都可以起类似的作用.阶为10的钟群隐藏在十进制数系之后，阶为24的钟群对应着24小时的钟，阶为60的钟群联系着时间的度量，阶为360的钟群与角度计算有关.

数学家将钟群命名为循环群，之所以这样称呼是因为这种群的元素在一个圈中循环变动（像一只钟）.例如三阶循环群的加法表为：

+	1	2	3
1	2	3	1
2	3	1	2
3	1	2	3

[118]

单　群

科学中任一分支的基本目标之一是确定并研究那些"基本对象"，其他所有对象可由它们构造出来.如在生物学中是细胞（或者是分子），化学中是原子，物理学中是基本粒子（目前是夸克）.在数学的很多分支中也是这样.典型的例子是数论（根据第1章描述的算术基本定理），素数是基本的建筑构件.上述例子中每种理论的基本对象在结构上都是单纯的，单纯的意思是指（从理论观点考虑）它们不能分成同类的更小的单元（原子不能用化学手段分裂，素数不能再用除法分解等）.

群论中的基本构件是单群.为了解释什么是单群以及任给一有限群如何能分解为组成它的各个单群，我们需要一个概念，即群的压缩像.大略地说，当我们构作一个群 G 的压缩像时，我们得到的是按比例缩减的 G 的一种变形. G 的群运算 $*$ 也反映到压缩像中，尽管只是某种约简形式.这有点像颠倒拿着望远镜观看一个物体：物体的主要特征被保存了，但是变小了，并且许多特征变得不再好区分.

稍微精确些说，如果从群 G 出发，构作一个 G 的压缩像 G'，你必须将 G 的每个元素 a 对应 G' 的一个元素 a'（称为 a 的像），使得 G'

的任一对元素 a,b 的像分别是 a',b' 的话,则 a' 与 b' 的乘积必须是 G 中元素 $a*b$ 的像(这样 G' 保存了 G 的结构).这里没有限制可能若干个 G 的元素对应 G' 中同一元素.通过这种"推倒"再"粘合"的过程使你从 G 得到 G',G' 的大小会缩减.数学家们称压缩像为同态像.

每个群 G 至少有两个压缩像,一个是 G 自身,即每个元素是它自己的像.这种情形显然满足作为压缩像的要求,这是一种极端情形;另一个压缩像是任一群 G 都有的另一种极端情形,它是 G 的点像,即只有一个元素即单位元 e 的群.注意,群的定义允许一个单位元自身构成一个群,显然它太平凡了.这种群的乘法表是

[119]

$$e * e = e.$$

在点像中,每个 G 的元素都有同一的像,即 e.所以这种情形也满足压缩像的要求.

钟群提供了一些例子,说明除了上面提到的两个平凡的群之外,存在着其他的压缩像,例如,设 G 是阶为 24 的钟群,设 G' 是阶为 12 的钟群.对 G 中从 1 到 12 的 n,其像 n' 就是 n 自己.从 13 到 24 的 n,$n' = n - 12$.于是 G' 是 G 的压缩像.例如取 G 中元素 7 和 18,它们的像分别是 7 和 6.在 G' 中 7 和 6 的和是 1.根据压缩像的定义,它应该等于 7 和 18 在 G 中的和的像.而在 G 中 7 与 18 的和是 1,1 在 G' 中的像恰是 1(注意,从 G 到 G' 的过程正是通常所用的从 24 小时的钟变为 12 小时的钟的方法).

注意在上面例子中很重要的一件事是 12(G' 的阶)除得尽 24(G 的阶).阶为素数的钟群没有不同于自身和点像的压缩像,这就给出了一族例子,它属于本章的中心概念:

单群是一种群,它的压缩像仅有其自身和点像.

通过压缩的方法,每个有限群可唯一地被分解为一些单群的集合,这与一个合数被分解为它的素数因子非常相似(见第 1 章).事实上还有更多的相似之处:每一个这样的单群分支的元素个数是原来

的群的元素个数的因子，而且所有这些数的乘积等于原来群的元素个数.但类似的东西到此为止.一方面，群的单群分支所包含的元素个数可以是合数(我们曾提到正十二面体的旋转对称构成一个60阶的单群).还有，在一个给定的素数集合中的所有素数的乘积是唯一的，而一个给定的单群集合常常可用不同的方法组合成完全不同的群. 〔120〕

分类问题

在完成了将单群视为有限群理论中的"基本粒子"之后，数学家们着手尝试单群的分类问题.大略地说，他们希望能够说出什么样的群是单群，什么样的群不是单群.自然，单纯性定义本身给出了一种答案：单群是那些仅有两个压缩像的群.但这不是数学家所要求的那种答案.他们希望的解答能够对作为群的单群的实际结构给出一种描述.这种描述可以是：给出一种一般的模式，适用于一族群；或是对个别的独特的群的结构的描述.

例如，一个显然的结论是所有素数阶钟群是单群(合数阶钟群则不是).事实上它们是仅有的交换单群的例子，因而我们已经得到了所有交换单群的完全分类(它们属于"正则"族).所以全部努力都花费在非交换群的分类上.大约从本世纪40年代以来，数学家开始在分类定理(虽然这不一定是他们心中的目的)方面工作，发现了几个无限单群的"正则族"，最后共发现了18个这种族，包括上述素数阶钟群族，以及另外一些容易描述的族(详见后文).同时也发现了几个高度不正则的独特的群，不能纳入任一已知的类型中.最早的5个奇怪的(后来所称的)散在单群，是于19世纪60年代被É·马蒂厄(Émile Mathieu)发现的，因此就称为马蒂厄群.最小的马蒂厄群的阶为7920，最大的阶为244823040.直到一个世纪以后的1965年，第六个散在群被Z·扬科(Zvonimir Janko)发现，称为扬科群，此群有175560个元素，每个元素都是7阶矩阵(群中的运算是矩阵乘法).

发现这个群的途径预示出这种方法将会导致最后发现26个散在单群.它来源于考查第十七个正则单群族,即1960年R·雷(Rimhak [121] Ree)发现的一族,现在称之为雷族.

正如我们就要解释的那样,每个单群对应着提供该单群的结构的信息的某些更小的群,称为对合的中心化子,其精确定义见后文.但雷群由二阶矩阵组成,矩阵中元素取自一个基数为3的奇次幂的有限数集.(若3的奇次幂为1,则这有限数集就恰好是1,2,3.)作为证明早期的局限形式的分类定理的尝试的一部分,必须证明雷群是具有下述性质的仅有的单群:与其对应的对合的中心化子由二阶矩阵组成,矩阵中的元素取自一个基数为某个素数p的奇次幂的有限数集中.显然此处的p应为3,最后证明了上述论断是正确的,除了一个例外,即p为5,奇次幂为1的情况.于是扬科就动手研究起这个例外情形.他的意图是扫除这个余留的障碍,即证明没有一个单群对应着包有5^1(即5)的集合.他没有达到这个目的,但却设法得到了更加奇妙的结果:他证明了如果存在这样一个单群的话,则该群必须有175560个元素.这样精确的结果预示着存在一个真实的群隐藏在这个背景中.经过大量的计算,扬科成功地找到了它.这是发现的第六个散在群.为了纪念扬科,它被命名为J_1.

对于雷族以外的单群族使用类似的技巧,扬科很快找到证据说明了还存在两个散在单群,其中一个有604800个元素,另一个有50232960个元素.但他未能具体找出这两个群.两个中较小的一个J_2最后由M·小霍尔(Marshall Hall Jr)和D·威尔士(David Wales)找到;较大的一个J_3由G·希格曼(Graham Higman)和J·麦凯(John McKay)找出(使用计算机进行演算).

通过或多或少相似的方法,这些年来人们陆续看到若干散在群被发现.到1980年,26个这种独特的群中最后一个被R·格里斯(Robert Griess)构作出来(有关它存在的猜想早在1973年就提出了).它是最大的散在群.实际上它得名为“大魔”.大魔的元素个数是:

808 017 424 794 512 875 886 459 904
961 710 757 005 754 368 000 000 000.

[122]

(粗略地说是8后面写上53个零.)它是阶为196883的矩阵的集合(矩阵的元素为复数).值得注意的是,格里斯是靠手算进行了确定"大魔"所需的全部计算的.这一事实说明这个群允许人用手算就能了解它,是充分跟人"合作的".于是格里斯给他的群起名为"友好的巨人".

大魔的发现是分类定理证明的最后过程中的一步.现在已经知道,有限单群包括18个正则无限群族(其中的第一个群族是素数阶的钟群),以及26个散在群,再没有其他的了.整个结果由500篇文章组成,在数学杂志上占用了几乎15000页.

十八个正则族和散在单群登场亮相

在数学的进步中,数学家常常做的事是阐述提出的定理并随后给出证明.而分类问题则与此大不一样:直到定理被证明之前,甚至连问题有多大也无法知道.例如也可能散在群多于26个,甚至可能有无限多个,这就意味着我们绝无可能达到目的.我们所做的大部分工作是在"让我们找出单群"的基础上进行的,而不是在奔向一个已明确叙述好的定理.这就很难确切地说出什么时候这方面的工作开始进入最后的分类工作.在1954年阿姆斯特丹的国际数学家大会上,R·布劳尔(Richard Brauer)在他的发言中提出了试图将单群(偶阶的,虽然后来证明这个限制是多余的)分类的一个方法,这可以算作一个起点.另一个较少争议的起点是1972年,那年,D·戈伦斯坦(Daniel Gorenstein)在芝加哥大学作了一系列的讲座,概括地提出了分16步走的纲领,它应该能导致分类问题的最终解决.在某种程度上看,最后的冲击是由1962年W·费特(Walter Feit)和J·汤普森(John Thompson)所得的关键性结果开始的,它也能看作是"结尾性工作的起点".无论如何,要想继续讲下去,我们首先必须谈谈出现在

[123]

完全的解答中的群的性质.

18个正则族中的第一个已经提到了,即:所有素数阶的钟群组成的族.第二个好像容易描述些.对任一大于4的整数 n,n 个符号的所有偶置换组成的群是单群,所有这样的群构成第二族.什么是 n 个符号的偶置换?考虑 $n=4$(第一个使人感兴趣的情形).取四个符号,如四个字母 A,B,C,D.按字母表顺序作成一个"字"$ABCD$.通过反复交换一对字母可将这四个字母排成 $4\times3\times2\times1=24$ 个不同的"字"(或序)中的任何一个.任一这样的重排叫做 $ABCD$ 的一个置换.通过偶数次交换得到的是偶置换,奇数次交换得到的是奇置换.如 $CBDA$ 是偶置换,从 $ABCD$ 出发,首先交换 A 和 C,然后交换 A 和 D 即得.$BACD$ 是一个奇置换,从 $ABCD$ 出发,经 A,B 这一对字母交换即得.

不是很妙吗?我们现在将上面做的与对称联系起来,而且不把置换看成是字母最后的排序,而看作是达到这一最后排序的一系列的对调.这意味着我们可将两个置换合并成为一个单个置换:如果 a 和 b 都是 $ABCD$ 的置换(即一系列对调),则 $a*b$ 也是一置换,它先实行对调 a,然后实行对调 b.例如,若 a 是先对调 A 与 C,再对调 C 与 D;又若 b 是对调 A 和 B,那么,从 $ABCD$ 出发,我们有:

a 将 $ABCD$ 变为 $DBAC$,

b 将 $DBAC$ 变为 $DABC$,

$a*b$ 将 $ABCD$ 变为 $DABC$.

显然运算 $*$ 满足结合律;单位置换 e 不改变任何东西,它的作用是恒等运算,即对任一个 a:

$$a*e=e*a=a.$$

任一置换的逆显然由同一个对换的反序作用而得到.于是,若 a 是先对调 A,C,然后对调 C,D,则 a^{-1} 为先对调 C,D,而后对调 A,C.[124](你自己检验 $a*a^{-1}=a^{-1}*a=e$.)这样,四个字母 A,B,C,D 的置换便组成一个群.偶置换自身也构成一群,它是全体置换群的子群,由全部置换中的一半组成(由两个偶数的和仍是偶数这一简单事

实可得出两个偶置换的乘积 * 是另一偶置换). 上述偶置换群称为 4 次交错群,它是一正四面体的旋转对称群的精确复制品(即两者有相同的乘法表,因而实质上是同一个群).

对 $n=4$ 的情形我们就讨论到此. 同样对任一大于 2 的 n,都可以得到 n 次交错群(当仅有两个符号时,只有一个非平凡置换,它是奇置换,故相应的交换群是一个元素的群). 当 $n=3$ 时,置换群有 $3\times2\times1=6$ 个元素,交错群是三阶钟群的镜面像(若 a 是 ABC 的偶置换,它先对调 A,B,然后对调 A,C,于是 a 将 ABC 变为 BCA,$a*a$ 将 ABC 变为 CAB,加上单位置换,这就是所有的偶置换,$(a*a)*a$ 将 ABC 仍变为 ABC,于是这架钟走了一圈).

对任一大于 4 的 n,n 次交错群是单群. 这是由于次数大于 4 的多项式方程(根式)不可解性所使然. 等一等,这不就是早些时候我们提到过的关于正十二面体的旋转群的单纯性吗? 而且由此导致伽罗华得到五次方程的有关结果. 确实,十二面体群恰是五次交错群.

在素数阶钟群和次数大于 4 的交错群之后,对另外十六个正则族要如此简捷地描述则难得多. 它们全是大小不一的矩阵群,有时这些族首先由它所包含的矩阵所描述;对另一些情况又须先定义其他术语,经过相当的努力之后才能得到其矩阵描述.

那么到底分类问题是如何解决的呢? 许多正则族在世纪之交时已经知道了,马蒂厄的五个散在单群就是那时发现的. 从观察发现,所有已知的非交换单群都含有偶数个元素,于是伯恩赛德(Burnside)猜想,这对于所有非交换单群都是对的,而不管它们总共有多少,也不管它们还可能有其他什么性质. 在 1962 年,伯恩赛德猜想被芝加哥大学的费特和汤普森所证明. 此结果获得 1965 年代数方面的科尔(Cole)奖. 费特-汤普森定理的证明占满了"太平洋数学杂志"整整 255 页的版面(像这样的数学杂志通常刊登 20 或 30 篇各种不同题目的论文),这预示了完全分类定理的最后证明将是极长的. [125]

随着费特-汤普森定理的出现,仿佛突然为沿着 1954 年布劳尔勾划的通向分类定理(前面提到过此事)的道路扫清了路障. 这个问

题有两个方面:一是要确定出单群(或单群族),这是分类所要求的.除了一个剩下的族和几个散在单群之外,这方面的工作在1960年就完成了(虽然当时对这点完全没有看清楚).另一方面是证明任一单群必定属于某个给定的范畴.这就使证明变得十分复杂.问题是你必须从完全任意的一个单群开始(即你所知的就是它是一单群),然后想方设法证明它是正则族的一个成员(或说是它的确切复制),或是所列出的散在群的一员.这就是布劳尔所建议的进攻路线.

他的思想是集中研究群中(非单位元的)a,它们满足 $a * a = e$.这样的元素称为对合.很容易证明任意含偶个元素的群至少包含一个对合.(你自己试一下.你需要知道的只是前面给出的群的定义.这个解法非常简捷精巧,是值得你花费些力气找到它的.)由费特-汤普森定理可得:每个非交换单群都包含对合.

布劳尔所作的事是计算几个已知的正规族里对合的中心化子(我们还记得,这是在费特-汤普森定理之前,但是在伯恩赛德提出他的猜想之后,他的猜想后来成为他们的定理).什么是中心化子?群 G 里的元素 g 的中心化子是群中所有使得 $a * g = g * a$ 的 a 所组成的集合.如果 G 是交换的,则任一元素的中心化子恰是 G 自身.自然,其他情形不必如此.可以相当直接证明的事实是:G 中任一元素的中心化子是 G 的子群.布劳尔的工作是鼓舞人心的.所有的对合的中心化子群具有像原来的单群同样的一般性结构,虽然是萌芽[126]状态的.这使他觉得,有可能从有关这些中心化子群的信息重造全部群,而且对某些特殊情形,他证实了他的估计.

布劳尔的工作不仅促使人们发现了很多散在群(前面已经提到三个扬科群),而且提供了使任意给定的单群纳入所提出的分类范畴的最初方法.首先证明,在给定的群中的对合的中心化子很像我们已知的分好类的一个单群的对合的中心化子,然后试图将这种高度局限的联系扩充为完全等价.这最后一步是没有容易方法的:对合的中心化子仅仅是全部对合中很小的一部分,所以这有一点像玩拼图游戏时想从一片拼图片导出全部图形一样.

在布劳尔之后进行的工作，属于戈伦斯坦1972年在芝加哥讲演中提出的16步纲领中的内容．戈伦斯坦本人认为，整个行动纲领可望在本世纪末实现．他的大部分听众则认为这未免太乐观了．但他们都没注意到听众中有一位刚刚完成了研究生学业的年轻数学家——加州理工学院的M·阿施巴赫尔(Michael Aschbacher)．他从一个被称为分支定理的关键性结果出发，以秋风扫落叶之势，一个接一个地证明出令人吃惊的成果．其结果是，所罗门能够以一项小的数学成果来标志该证明的最终完成；那是1980年的事，恰好在戈伦斯坦的讲演发表8年之后．(阿施巴赫尔则因其出色的工作而获1980年代数方面的科尔奖．)

沿着这条路走，其余的各种散在群被发现了．正规族除了前两个以外全部是由某种矩阵集合组成的，有时还需求助于计算机来计算．由于这26个群在所有无限个单群中占极少数，显然它们是非常特别的，因此当我们得知它们与别的数学分支有联系时是一点也不奇怪的．例如，1968年剑桥大学的J·康威(John Conway)发现的三个散在单群，现在以他的名字命名，它们是奠基于利奇格(Leech's lattice)的，后者是一种数学结构，它起源于设计纠错码(这是一种信息的加密传输方法，使得失真和随机错误可以得到补偿)．有两个马蒂厄散在单群是与常用于军事目的的戈莱(Golay)纠错码相关．这一类的联系说明了对分类定理有兴趣的某种原因，但分类定理在群论之外的最主要的"名声"无疑来自于其证明之令人难以置信的长度．本章的 [127] 最后几句话我要留给M·阿施巴赫尔，他在得到最后的证明中起了十分大的作用．在1980年完成整个证明后不久，他曾回顾了这个证明．他写道：

"这里所涉及的数学的大部分内容是最近得到的．无疑，一旦有时间深思熟虑这些技巧，它们将会得到改进．然而还是很难想象这个定理会有一个短的证明．我个人很怀疑将来会出现任何一类短证明．

长的证明会使很多数学家感到困惑．一方面，当证明的长度增加时，错误的概率也增加了．在分类定理证明中出现错误的概率实际上是1．另一方面，任何单个的错误不能被容易地改正的概率是0．由于该证明是有限的，所以定理是错误的概率接近于零．随着时间的推移，我们有机会推敲证明，对它的信任程度必定会增加．

现在也许应该考虑以下的可能性：有一些自然的基本的定理，它可以简明地叙述出来，但没有简短的证明．我猜想分类定理就是这类成果．当我们的数学变得更成熟时，我们可能更会经常地碰到这类定理．”

阅读文献

关于群论的比较好懂的引言可在 Ian Stewart 的 Concepts of Modern Mathematics(Pelican, 1981)一书的第7章中找到．

程度高一些的读物，涉及到本章课题的细节，可参看 Daniel Gorenstein 的 Finite Simple Groups (Plenum, 1982)．

有限单群分类的工作的权威性描述可参见 Daniel Gorenstein 的两卷本的 [128] The Classification of Finite Groups (Plenum, 1982)．

（袁向东、冯绪宁译）

第6章 希尔伯特第十问题

历史回顾

1900年8月，全世界最优秀的数学家云集巴黎，出席第二次国际数学家大会(这样的大会每隔四年在世界上不同的地点举行一次，除战争时期外没有中断). 他们中有一位就是38岁的哥廷根大学教授大卫·希尔伯特(David Hilbert). 作为当时的领头数学家之一，希尔伯特应邀向大会作主要报告，报告日期是8月8日.

由于此次大会的召开正好赶上20世纪的头一年(其实是为此而特意将会期提前了一年)，希尔伯特决定利用这一机会为未来的发展指引方向，而不是(像通常这类报告所做的那样)简单地回顾某些最近的工作.

希尔伯特大声疾呼："在我们中间，常常听到这样的召唤：这里有一个数学问题，去找出它的答案！你能通过纯思维找到它. 因为在数学中没有 ignorabimus(不可知). "为了强调这种召唤，希尔伯特在会上提出了不是一个而是二十三个尚未解决的重要问题——其中任何一个问题一旦获解，都将标志着数学知识的重大进展. 这些问题多数是(或者后来变得)以特殊的名字而著称，如连续统问题(在希尔伯特的表中名列第一，参阅第2章)，或黎曼问题(参阅第9章)等. 但有一个问题却特别地以其在希尔伯特表中的排序——第十而变得众所皆知.

[129]

希尔伯特第十问题来源于一本名叫《算术》(Arithmetica)的代数

著作,该书写于公元250年左右,作者是亚历山大城的丢番图(Diophantus)(参见第8章).根据这本著作中所考虑的问题的类型,今天的数学家们就用"丢番图方程"这个名称来表示那些有一个或几个变元的整系数方程,它们的求解仅仅在整数范围内进行.正是最后这个限制使丢番图方程的数学与实数(也可能为复数)范围的方程求解有根本的不同.(事实上"丢番图方程"这一术语初听起来真有些令人混淆."丢番图"作为形容词主要不是修饰方程,而是用来说明所寻求的解的类型.因此方程

$$3x^2-5y^2+2xy=0$$

如果是在实数范围内求解,就简单地称之为"方程",而如果只要求整数解,则称它为"丢番图方程".)

解丢番图方程与在实数范围内解同样的方程很不一样.例如,设有方程

$$x^2+y^2=2, \tag{5}$$

如果将它看成是求实数解的通常方程,那么就存在着无穷多个解.在 $-\sqrt{2}$ 与 $+\sqrt{2}$ 之间给定任一实数 r,如取

$$s=+\sqrt{2-r^2},$$

则 $x=r, y=s$ 就是一个解.但如果将它看成是一个丢番图方程,那就只能得到四个解:

$$x=+1, y=+1; x=+1, y=-1;$$
$$x=-1, y=+1; x=-1, y=-1.$$

如果将方程稍微变动一下,比如说变成

$$x^2+y^2=3, \tag{6}$$

那么仍然可以得到无穷多个实数解,但却根本不存在整数解了.作为[130]丢番图方程,方程(6)不可解.那么方程(5)与(6)究竟有什么区别?更一般地说,是否有一种可以判别任意丢番图方程可解性的方法?例如,能不能编写出一套计算机程序,对于任给的丢番图方程,它都可以告诉你该方程是否有解?这实质上就是希尔伯特第十问题所问的内容.在前人所做大量工作的基础上,这问题到1970年才被俄国

数学家尤里·马蒂雅舍维奇(Yuri Matyasevich)解决，而这方面的努力可以追溯到1930年代，并包容着数理逻辑、计算理论和代数学等多方面的结果.

丢番图方程和欧几里得算法

最简单的丢番图方程是只有一个未知数的线性方程.事实上我们关于丢番图的生平所知道的唯一信息就是从一个这样的方程得到的.有一个4世纪的数学问题说：丢番图的童年占其一生的六分之一，又过了一生的十二分之一他开始长胡须，再过七分之一他结了婚，婚后五年生子，儿子的寿命是父亲的一半，且比父亲早亡四年.如以 x 表示丢番图的寿命，上述信息就给出方程

$$\frac{1}{6}x+\frac{1}{12}x+\frac{1}{7}x+5+\frac{1}{2}x+4=x,$$

解出 $x=84$.(严格地说这并不是一个丢番图方程，因为其系数非整数，但若用系数各分母的最小公倍数遍乘各项，就可以得出一个整系数方程.)不管丢番图是否真活了84岁，只有一个未知数的线性丢番图方程的求解总是一件轻而易举的事情.方程

$$ax=b$$

有整数解的充要条件是 a 整除 b，这时解就等于 b/a.这条件非常简单，因此很容易写出一个计算机程序，我们可以通过它很快地判断这样一个丢番图方程是否有解. [131]

有两个未知数的线性丢番图方程情形又怎样呢？这时同样有一种简单的方法可以判断方程是否有解.为了判断方程

$$ax+by=c$$

是否有整数解，可以首先计算 a 和 b 的最大公因数，比如说是 d.如果 d 整除 c，方程就有整数解；如果 d 不能整除 c，方程就无整数解.

例如方程

$$6x + 15y = 12$$

有没有整数解？6和15的最大公因数是3,3可以整除12,因此方程有整数解(例如 $x = 7, y = -2$ 就是方程的解.)

请注意,对于给定的丢番图方程来说,判别它有没有解和求出它的解是性质完全不同的问题.有可能解的存在性很容易判断,但要实际求出一个解却极为困难.(虽然,如果你能够求出一个解来,那就同时知道了它的存在！能够被找到的东西一定存在,反之,存在的东西却未必能被找到.)对于两个未知数的线性丢番图方程来说,不仅有一种简单的判断解的存在性的方法,而且还有一套具体求解的机械化程序.大多数初等数论课本对此都有详细介绍①.解法的关键是下述求最大公约数的欧几里得算法.

已知两数 x 和 y,以 $x \bmod y$ 表示 x 被 y 相除所得余数(见第1章).两个已知数 a 和 $b(a > b)$ 的最大公约数可以计算如下.设 $a \bmod b = r_1$,再设 $b \bmod r_1 = r_2$, $r_1 \bmod r_2 = r_3$.如此继续下去,直到余数为零:$r_{n-1} \bmod r_n = 0$.这时 r_n 就是 a 和 b 的最大公约数.

[132] 例如求133和56的最大公约数,可以进行如下:

$$133 \bmod 56 = 21,$$
$$56 \bmod 21 = 14,$$
$$21 \bmod 14 = 7,$$
$$14 \bmod 7 = 0.$$

因此133和56的最大公约数是7,即最后一个非零余数.(现在你也许愿意亲自验算一下81和25的最大公约数是1.)

上述方法最早见于公元前350～300年左右写成的欧几里得《原本》第Ⅶ卷,这就是为什么它被称作欧几里得算法的原因.然而确切地说,什么是"算法"？对于希尔伯特第十问题来说,这可是关键的问题.在试图回答这一问题之前,让我们先简要回顾一下,迄至希尔伯特讲演之时,人们对于丢番图方程的求解还知道些什么.

---

① 如《初等数论》,David Burton 著(Allyn and Bacon,1980).——原注.

事实上所知甚少(至今仍然如此). 两个以上未知数的线性方程可以用一种推广的欧几里得算法来处理，即上述两个变量情形的推广. 对于有一个或二个未知数的二次方程，如

$$x^2 - 3x + 4 = 0$$

或

$$3x^2 - 5xy + y^2 = 7,$$

高斯的一种重要理论提供了判断给定方程是否有解的方法.(这就是著名的二次互反律，第 63 页上已有介绍.)但是，除了对个别特殊情形可以使用一些巧妙的方法以外，上面所说的大概就是我们所知道的一切了.(一个特别重要的“特殊情形”涉及丢番图方程

$$x^n + y^n = z^n,$$

这里 n 至少等于 2. 当 $n > 2$ 时这方程解的存在性就是著名的费马最后定理问题，我们将在第 8 章中作详细介绍.)

现在我们就来讨论什么是“算法”的概念? [133]

算法与图林机

公元 825 年左右，一位名叫阿尔·花拉子米(al-Khowarizmi)的波斯数学家写了一本书，书中概括了进行数字四则算术运算的法则，所有的数字都是用今天的印度十进制形式来表示的(按个、十、百位等排列，并有表示分数量的小数点). 现代名词“算法”(algorithm)就来源于这位数学家的名字.

所谓算法就是逐步(step-by-step)执行某类计算的方法. 是否能明确写出或具体说明运算指令并不重要. 重要的是这些指令应当是完全的，并且没有二义性，没有随意选择的余地，同时还应对所有的初始数据而不仅是某些特殊数值有效. 上节所述欧几里得算法是一个很好的例子. 算法指令明确告诉你每一步该做什么，同时该方法对所有的数 a 和 b 都适用(a 大于 b——为了适用于所有的情形，只需简单地加一道初始指令：按大数在前小数在后的顺序写出这两个数

即可).算法的其他例子有花拉子米在他的著作中制定的十进数的加、减、乘、除法则.

希尔伯特第十问题是问:是否存在一种可以判别任给的丢番图方程有解的算法?对于某些特别简单的丢番图方程,这样的算法是存在的.如前所述,对于线性方程和至多两个未知数的二次方程,这样的算法确实存在.但是否存在一个能适用于所有情形的算法?如果答案是肯定的,那么为了证明你的结论,只要将这个普适的算法写出来就行.但是假定答案是否定的,那又怎样来证明你的结论呢?"逐步指令集"的概念对于确定具体的算法来说是适用的,但它总的来说过于含糊,不能用来证明执行某一特定任务的算法之不*存在*.因此就需要给算法下一个更为严格的数学定义.

有了今天的计算机技术,人们可以将"算法"定义为对特定的机器用特定的计算机语言编写的一套计算机程序.这样的定义当然是很精确.但同时也产生一些明显的问题.首先是什么样的语言和什么样的机器?还有对机器所能处理的数字的大小和可以利用的存储量有何限制?后来终于弄清了,假如你准备通过取消对数据大小的所有限制而使问题理想化,那么语言和机器的选择对于"算法"的最终定义来说是无关紧要的.不管用什么语言和机器,所产生的可计算函数集都相同:一种计算能够在一种机器上用一种语言进行,当且仅当它能在任何其他机器上用任何其他语言进行.这在直觉上是很清楚的,因为在最基本的运算层次上,计算机所做的一切就是处理0和1.

[134]

事实上,为了得到可行的"算法"定义,并没有必要求助于计算机技术.早在计算机时代来临前的1930年代,数理逻辑学家(他们中最重要的有E·波斯特、A·丘奇、S·克林、K·哥德尔和A·图林)已经给出过若干种定义.这些定义采用了很不相同的途径:"方程演算"(equational calculus),"递归函数"(recursive functions)演算,以及各种抽象"计算机".不过在每一种情形,最终得到的"可执行计算"的概念都一样,因此就"算法"概念的定义而言,你可以选择其中的任何一种

途径.我们不妨就选择最简单的方法,这种方法是由英国逻辑学家艾伦·马西森·图林(Alan Mathison Turing)最先提出的.

图林假设存在一种抽象的计算机,这种计算机今天就叫图林机.它由一个读写头和一条无限长的带子组成,带子可以双向通过读写头,带子上分成许多小格(见图 30).每个格子可以是空白的,也可以记有一个符号,这符号是取自固定的符号系统(0 和 1 两个符号就够了,但在整个运算中符号系统的确切选择并不重要).在任一时刻,读写头将处于有限多个不同状态(State)中的一个状态.(两个状态就够了,但具体的状态数不是重要因素.) 机器的运作是按逐步进行的方

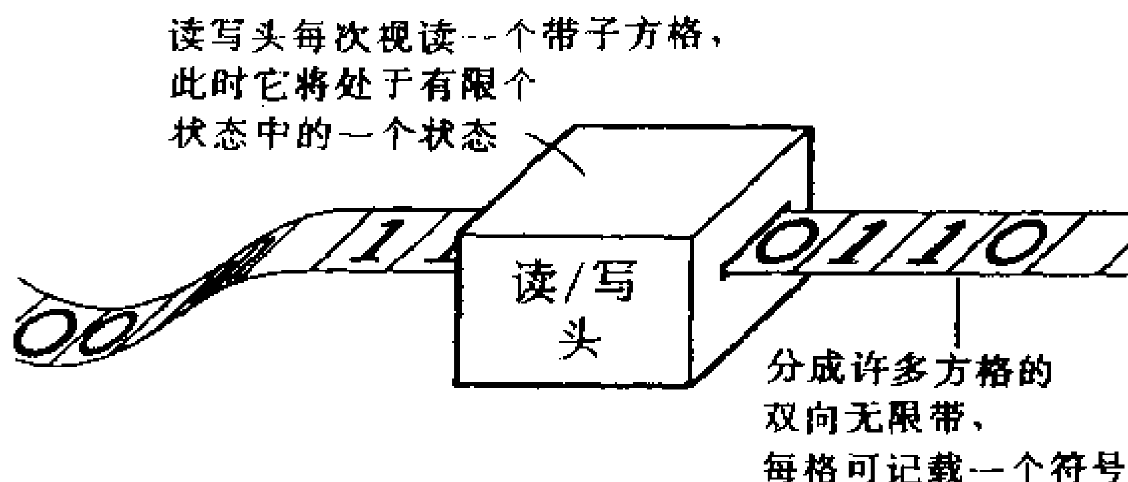


图 30 图林机.这种假想的计算机是由英国数学家艾伦·图林在 30 年代发明的,目的是为计算的研究提供抽象的理论框架.图林机虽然很简单,但可以证明,任何计算不管有多复杂,都可以用图林机来完成.图林机概念使人们能给"算法"以精确的定义,即把算法看作为图林机程序.

读写头在任一时刻必处于有限个确定状态中的一个状态.机器的动作是按离散的步骤进行.每一步的具体动作由读写头当时状态和它正在视读的带子方格内容所决定.机器的运作由一套程序来控制,程序由一张表格组成,指明了跟随着每一组由读写头状态和带子输入符号决定的信息之后机器应该进行什么动作(参阅框图 B).

式,每一步则由三个不同的动作组成.在任何一个时刻,读写头将视读带上的一个方格.它的行为由该方格上的内容和机器的状态所决定.根据这两个因素,机器抹去带上原有的符号;然后或者使该方格保持空白,或者则写上另外的(也可能是相同的)符号;然后让带子通过读写头,朝两个方向之一移动一个方格,最后机器进入了另一个(也可能是相同的)状态.机器的行为从头至尾是由一个指令集(instruction set)所决定,这个指令集明确地告诉你——对于每个状态和每一次可能的符号识读——应该执行哪三个动作.初始数据(如果有的话)也是写在带子上(根据某个符号系统,其具体选择并不重要),[135] 机器的运作从读写头视读第一个方格数据开始.一旦计算结束,机器就进入一个特别的停止状态.运算过程所产生的任何结果都记录在带子上,可以从机器停止时所视读的那个方格开始去寻找这些结果.

利用图林机,算法可以被定义为一串指令,它们按上述的方式来决定机器的行为.很明显,用如此简单的机器即使进行最基本的运算,所需的"算法"也将十分繁琐(参阅框图 B).但这一概念的意义,主要是在于它提供了算法(以及计算)的精确定义,这定义非常简单,能够进行数学处理,并且适于执行任何"算法计算".概念本身并不要求实际制造出这样一台机器来——虽然已有许多热心人做了大量的 [136] 尝试!

框图 B:一个简单的图林机程序

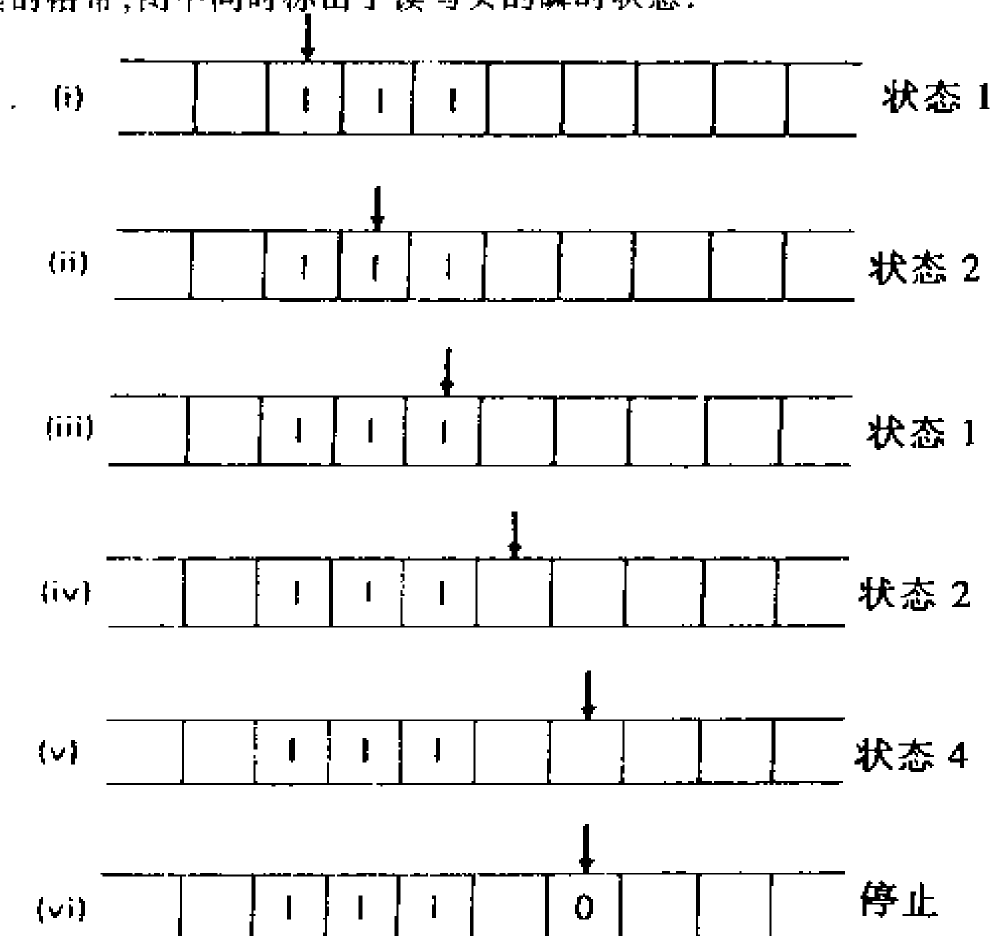
在本例中,机带符号表仅由 0 和 1 组成.正整数用 1 的连续序列来表示,n 个 1 就表示数 n(这样正整数被记作:1,11,111,1111,…).总共有五个状态,分别记为 1,2,3,4 和 H(特别停止状态).程序的对象是要确定输入带上的一个给定的整数是偶数还是奇数.如果是偶数,机器就应输出一个 1 并停机;如果是奇数,机器应输出一个 0 并停机.这个输出的数据出现在带子上表示整数的数据之后,二者隔一空格.假设输入的整数这样排列,使得读写头从它的第一个数字开始自左向右地视读.

在下表所列程序中,b 表示空格,R 表示向右视读下一个格带.更复杂的程序可能会出现向左的移动.

条件		动作		
状态	输入	输出	状态	移动
1	1	1	2	R
1	b	b	3	R
2	1	1	1	R
2	b	b	4	R
3	-	1	H	-
4	-	0	H	-

[137]

下图逐步列出输入数据 3(在带子上即 111)的程序动作. 箭号指示正被视读的格带,图中同时标出了读写头的瞬时状态.



你也许会愿意自己来试试,列出输入一、二个其他数据的程序动作,或者编写出执行其他简单任务的程序,例如用上述方式表示的两个正整数的加法程序.

[138]

可 计 算 集

对于解决希尔伯特第十问题来说，一个关键的概念是整数的可计算集(computable set). 所谓可计算集是指一个整数的集合 S，对于这个集合来说，存在着能判定一个数是否属于 S 的机械化(或算法化)方法. 用图林机的术语来说，一个整数集被称为可计算的，如果存在着一个图林机程序，按照此程序，当给定任一整数作为输入后，若该整数属于 S，那么机器将输出数据 1 并进入停止状态，若该整数不属于 S，则机器将输出数据 0 并进入停止状态. 例如，所有偶数的集合 S 是可计算集. 框图 B 所列程序可以完成必要的运算. (这程序实际上只能处理正整数情形. 为了允许出现负数的情形，则需要对正负数的编码作出某种规定，例如以第一个符号表示正负号. 作为练习，你不妨可以试试修改一下框图 B 中所给的程序，以便能够处理上述更一般的情形.)

注意在上述可计算集的定义中，对于每次输入，图林机程序都会产生一个结果——它决不会进入无限的循环或开始对某种根本不存在的数据进行无休止的搜索，而这样的情况在实际的计算机上用实际程序进行计算时是常会发生的. 一个允许有这种"无尽计算"的较弱的概念是所谓整数的可列举集(listable set)(数学家们称之为递归可数集(recursively enumerable set). 用图林机的术语来说，一个整数集 S 被称为可列举的，如果存在着一个图林机程序，按照此程序，若给定任一整数作为输入后，当且仅当该整数属于 S 时，机器输出数据 1 并进入停止状态. 当输入整数不属于 S 时，机器可能输出数据 0 并进入停止状态，也可能根本不停止. 这样，当你用该程序对给定的输入整数 N 进行计算时，如果刚好 N 是 S 的元素，那么此程序最终将会帮你得到这一结论，但如果 N 不是 S 的元素，那么你可能永远也不会得到这样的结论——计算可能会永远不停地进行下去，尽管你决不相信它不会停止. 因此，这是一种非常片面的形势.

正如你可以想象的那样，可计算集与可列举集这两个概念之间存在着密切的关系．一个整数集是可计算集，当且仅当 S 和 $\bar{S}$ 都是可列举集，这里 $\bar{S}$ 是 S 的补集（即所有不属于 S 的整数的集合）．这一点很容易证明．如果 S 是可计算集，那么任何可以检验 S 的可计算性的程序 P 也将能用来检验 S 的可列举性，而为了得到一个能证 [139] 明 S 是可列举集的程序，只要取上述的程序 P，并将其最后一步中的输出 0 改为 1，输出 1 改为 0 就行了．反之，如果 S 和 $\bar{S}$ 都是可列举集，那么为了得到一个能证明 S 是可计算集的程序，做法如下：设 P 和 Q 分别为给出 S 和 $\bar{S}$ 的可列举性的程序．如果有两台图林机，其中一台执行程序 P，另一台执行程序 Q，当你同时将给定整数 N 输入这两台机器，若 N 属于 S，则程序 P 最终将输出 1 并进入停止状态，而若 N 属于 $\bar{S}$，那么程序 Q 最终将输出 1 并进入停止状态．因此这两台机器合在一起就可以给出一种机械化方法，来判定一个给定的整数是否属于 S．这在直观上已经说明了 S 是可计算集．为了根据定义来精确地证明这一点，必须构造出一个能同时完成程序 P 和 Q 的任务的图林机程序．一个显而易见的做法是编写一个程序 R，它将对给定的整数输入轮流地执行 P 和 Q（比如说各走 100 步），直到其中的一个输出数据 1 并进入停止状态．如果是 P 做到了这一点，就确定 R 输出 1 并进入停止状态，如果是 Q 先做到这一点，就确定 R 输出 0 并进入停止状态．显然，程序 R 就验证了 S 的可计算性．

以上是关于可计算集与可列举集的相互关系所能得到的最好结果了．这两个概念肯定是不一样的．每个可计算集当然也是可列举集，但却存在着非可计算集的可列举集．为了构造这种集合的例子，需要引进“通用”(universal)图林机的概念——即可以模拟任意图林机程序的图林机程序．图林证明了这样的程序是可以编出来的．虽然构造这种程序的具体细节技术性很强，但其基本思想却很简单．首先列出所有的图林机程序：$P_1, P_2, P_3, \cdots$．通用程序执行如下：首先在机带上输入一个自然数 N，只要通用程序读到 N，接着就执行程序 P_N（在 N 之后所需要的任何数据都可在带上找到）．

现在我们就来构造一个非可计算集的可列举集 S. 设 S 是所有这样的自然数 N 的集合, 使得程序 P_N 对输入 N 执行结果是输出1并进入停止状态(此处 $P_1, P_2, P_3, \cdots$ 是为了得到上述通用图林机程序而列出的所有的图林机程序). 用通用图林机程序很容易证明 S 是可列举集: 只要在通用程序开头的地方加上一个小程序: 在机带上输入一任意整数并重复一次, 然后让读写头从这两个重复数据的第一个开始视读. (自己检验一下证明过程对读者来说是一个很好的练习.)

[140]

为了证明 S 是非可计算集, 我们假设结论相反, 即 S 是可计算集. 那么如前所述, $\bar{S}$ 将是可列举集, 因此就存在着一个可以证明这事实的程序. 这个程序必定出现在所有程序组成的表 $P_1, P_2, P_3, \cdots$ 中, 比如是 P_k. k 或者属于 S, 或者不属于 S, 二者必居其一. 如果 k 属于 S, 则 k 不属于 $\bar{S}$, 因为 P_k"列举"$\bar{S}$, 故从 k 出发 P_k 不可能得到输出 1 并进入停止状态的结果. 因此 k 不满足定义 S 的条件, 也就是说不属于 S. 这样, 如果 k 属于 S, 则可推出它不属于 S! 另一方面, 如果我们假设 k 不属于 S, 则会发生什么情况呢? 这时 k 必属于 $\bar{S}$, 故对 k, P_k 将最终输出 1 并进入停止状态, 于是 k 满足 S 的定义条件, 这就是说 k 属于 S. 这样, 如果 k 不属于 S, 则可推出它属于 S! (对于读者来说, 这种情况似曾相识. 请参阅第 2 章, 特别是罗素悖论和康托定理的证明.) 这样我们就得出了一个矛盾, 唯一正确的结论是: S 不是可计算集, 正如一开始所假设的那样.

有了最后的这个结果, 我们现在可以来介绍希尔伯特第十问题的解了.

希尔伯特第十问题

希尔伯特实际上并没有直截了当地问是否存在可判定已给丢番图方程有没有解的算法. 他的问题是要求构造这样的一种算法. 以下的话是引自他本人的演说:

> “给定了一个有任意个未知数的、系数为有理整数的丢番图方程，试设计一种方法，根据这种方法可以通过有限步运算来判别该方程是否有有理整数解.”

另一方面，他在其演讲的其他地方说道（关于一般的问题）：

> “有时会碰到这样的情况：我们是在不充分的前提下或不正确的意义上寻求问题的解答，因此不能获得成功. 于是就会产生这样的任务：证明在所给的前提和所考虑的意义下原来的问题是不可能解决的.”

对于希尔伯特第十问题来说，情况正是如此. 马蒂雅舍维奇在 1970 年恰恰证明了希尔伯特要求的算法并不存在. [141]

这方面第一个认真的尝试是由戴维斯（Martin Davis）在 1950 年进行的. 他的方法是这样的（如果觉得难懂，可参看后面的例子）：证明对每个可列举集 S，都有一个相应的整系数多项式 $P_S(x, y_1, y_2, \cdots, y_n)$，使得正整数 k 属于 S 当且仅当丢番图方程

$$P_S(k, y_1, y_2, \cdots, y_n) = 0$$

有一个解.（P_S 的次数和它所含有的变元个数并不重要. 当这问题最终按戴维斯所建议的思路被解决后，人们发现 P_S 的次数不必大于 4，而 n 不必大于 14.）

例如设 S 是所有这样的正整数的集合，它们不能对某个 k 表示成 $4k+2$ 的形式. 这样

$$S = \{1, 3, 4, 5, 7, 8, 9, 11, \cdots\}.$$

实际上 S 正好是所有能表示成平方差形式的整数集（即所有形如 $a^2 - b^2$ 的数集）. 这样

$$1 = 1^2 - 0^2, \quad 3 = 2^2 - 1^2, \quad 4 = 2^2 - 0^2, \quad 5 = 3^2 - 2^2,$$

但却不存在数 a 和 b 使得

$$6 = a^2 - b^2.$$

一般的证明可以进行如下：如果 n 属于 S，则它必可表成下列形式之一：$4k$，$4k+1$，或 $4k+3$. 在第一种情形有

$$n=\left(\frac{n}{4}+1\right)^2-\left(\frac{n}{4}-1\right)^2.$$

在其他两种情形则有

$$n=\left(\frac{n+1}{2}\right)^2-\left(\frac{n-1}{2}\right)^2.$$

[142] 另一方面，每项平方或者是 4 的倍数，或者是 4 的倍数加 1，这取决于它究竟是偶数的平方还是奇数的平方. 因此两个平方的差决不可能是 4 的倍数加 2，因此不属于 S 的数不可能表示成两个平方的差.

现在假如我们把集合 S（显然是可列举集）与多项式

$$P_S(x,y_1,y_2)=y_1^2-y_2^2-x$$

联系起来. 那么不难验证：一个正整数 k 属于 S 的充要条件是丢番图方程

$$P_S(k,y_1,y_2)=0$$

有解；也就是说当且仅当方程

$$y_1^2-y_2^2-k=0$$

有一个整数解.

当然，上述例子之所以成立，是因为已经提到的 S 的特殊性质. 戴维斯想做的事情是使每一个可列举集 S 都与某个适当的多项式 P_S 相联系. 由此可以推出希尔伯特第十问题所要求的那类算法并不存在. 为了证明这一点，我们假设结论相反，即存在着这样的算法. 设 S 是在上节中所构造的非可计算性的可列举整数集. 根据我们的假设，存在着能判定丢番图方程可解性的算法，因此必有一个图林机程序 H，若丢番图方程

$$P_S(k,y_1,y_2,\cdots,y_n)=0$$

有解，则它对输入 k 运算的结果是输出 1 并进入停止状态；而若上述丢番图方程无解，则 H 对输入 k 运算的结果是输出 0 并进入停止状态. 但由于 S 和 P_S 之间的关系，H 就验证了 S 是可计算集，这与

S 的选定相矛盾.因此这样的程序 H 不可能存在.换句话说,不存在希尔伯特所要求的那种算法.

遗憾的是,尽管这一方法原则上可行,但戴维斯却没有能证明这样的多项式 P_S 一定存在.后来发现解决问题的钥匙包含在由朱莉亚·罗宾逊(Julia Robinson)开创的一些工作中.罗宾逊研究了那些可以用丢番图方程来定义的集合,并且发展了各种技巧来处理其解按 [143] 指数形式增长的方程.1960 年她与戴维斯和普特南(Hilary Putnam)合作,证明了:只要能够找到一个丢番图方程,其解在一定的意义上呈现指数式增长,那么就可以按戴维斯设想的方式用丢番图方程来刻划每一个可列举集,因此就解答了希尔伯特第十问题.但他们也并未能找到这样的一个方程,因而长期停滞不前.直到十年以后,尤里·马蒂雅舍维奇终于奏响了凯歌,马蒂雅舍维奇成功的地方,正是三位美国数学家失败之处.他的工作利用了一个著名的数列,这数列来源于一个 12 世纪的、与兔子有关的数学问题.

斐波那契的兔子与马蒂雅舍维奇的解答

1202 年,意大利数学家、比萨的列奥纳多(Leonardo)出版了他的著作《算盘书》(Liber Abaci)(该书以"斐波那契"——Fibonacci 的名义出版,Fibonacci 源于拉丁文 filius Bonacci,意思是"Bonacci 的儿子").这是一部有影响的著作,它将印度-阿拉伯十进数系引进了西欧.该书讨论的问题中有一个是这样的:

> 某人在一处有围墙的地方养了一对兔子.问这对兔子在一年的时间内总共生育了多少对兔子?假定每对兔子每个月生育一对兔子,新兔从第二个月开始生育.

这里假定了:最初的一对兔子经过一个月后才开始生育;兔群中没有死亡;每对兔子有规则地连续生育.不难看出,按月份列出的成

年兔子数形成如下数列：

$$1,1,2,3,5,8,13,21,34,\cdots,$$

这个数列产生的规则很简单，即开头两个数1以后的每个数都是由在前面的两个数相加而得. 这样，$2=1+1$，$3=1+2$，$5=2+3$，$8=3+5$，等等.

[144]

后来人们发现这个简单的数列本身具有某些有趣的性质，同时还有一些惊人的应用.（例如在计算机数据库理论中，在欧几里得算法的计算有效性研究中都用到了这一数列.）就希尔伯特第十问题而言，斐波那契数列的重要性是在于这样的事实，即它呈现出指数式增长. 数列中的第 n 个数近似等于

$$\frac{1}{\sqrt{5}}\left[\frac{1}{2}(1+\sqrt{5})\right]^n.$$

（n 越大，逼近度越高.）这就是说，利用前面提到的戴维斯－罗宾逊－普特南的结果，为了解决希尔伯特第十问题，只要能找到一个丢番图方程，其解与斐波那契数适当相关就行. 这正是马蒂雅舍维奇所做的事情. 为了得到他所发现的丢番图方程，读者可以从下列的十个多项式方程开始：

$$\begin{aligned}
u+w-v-2&=0,\\
l-2v-2a-1&=0,\\
l^2-lz-z^2-1&=0,\\
g-bl^2&=0,\\
g^2-gh-h^2-1&=0,\\
m-c(2h+g)-3&=0,\\
m-fl-2&=0,\\
x^2-mxy+y^2-1&=0,\\
(d-1)l+u-x-1&=0,\\
x-v-(2h+g)(l-1)&=0.
\end{aligned}$$

在这些方程中 u 和 v 的值有如下的关系：即 v 是第 $2u$ 个斐波那契数，这就足以满足戴维斯－罗宾逊－普特南结果的要求了. 现在你只

要简单地将这十个方程每一个都进行平方并把它们加在一起得出一个大方程来，这个方程就是解决希尔伯特第十问题所需要的方程. 当然你还可以得到更多的东西.

从希尔伯特问题来看，戴维斯－罗宾逊－普特南－马蒂雅舍维奇给出的是否定的解答：它证明了所要求的算法并不存在. 但实际上这却是一个非常肯定的数学结果. 根据这一结果，每一个可列举整数集都可以用一个丢番图方程来刻划：如果 S 是一个可列举集，则必有一个整系数多项式 $P(x,y_1,y_2,\cdots,y_n)$，使得数 k 属于 S 的充要条件是丢番图方程

$$P(k,y_1,y_2,\cdots,y_n)=0$$

有解.

例如，素数集是一个可列举集.(事实上它是一个可计算集. 编写一个可以进行素性检验的计算机程序并不困难，虽然正如第 1 章中所指出的那样，要编写一个有效的素性检验程序却不那么容易.)因此素数集可以用一个丢番图方程来描述. 运用一点代数技巧就可以证明：存在着一个多项式 $P(x_1,\cdots,x_n)$，其正值(当变元 $x_1,\cdots,x_n$ 取遍所有的整数)恰好是素数全体. 此结果解决了一个长期悬而未决的难题，即素数是否可以作为多项式函数的值来得到.(虽然要注意并不是这函数的所有取值都是素数——它同时还产生负值，这些负值可能是也可能不是负素数. 不过其正值包括了所有的素数，同时也不可能出现其他的正值.)

遗憾的是，马蒂雅舍维奇的结果只是说明这样的素数生成多项式一定存在. 它并没有指出怎样具体构造这样的多项式. 经过大量艰巨的努力，琼斯(James Jones)、萨托(Daihachiro Sato)、瓦达(Hideo Wada)和威恩斯(Douglas Wiens)等人终于在 1977 年找到了一个这样的多项式. 这是一个有 26 个变元的 25 次多项式：

$$\begin{aligned}&(k+2)\{1-[wz+h+j-q]^2\\&-[(gk+2g+k+1)(h+j)+h-z]^2\\&-[2n+p+q+z-e]^2\end{aligned}$$

$$
\begin{aligned}
&-[16(k+1)^3(k+2)(n+1)^2+1-f^2]^2\\
&-[e^3(e+2)(a+1)^2+1-o^2]^2\\
&-[((a^2-1)y^2+1-x^2]^2-[16r^2y^4(a^2-1)+1-u^2]^2\\
&-[((a+u^2(u^2-a))^2-1)(n+4dy)^2+1-(x+cu)^2]^2\\
&-[n+l+v-y]^2-[(a^2-1)l^2+1-m^2]^2\\
&-[ai+k+1+l-i]^2\\
&-[p+l(a-n-1)+b(2an+2a-n^2-2n-2)-m]^2\\
&-[q+y(a-p-1)+s(2ap+2a-p^2-2p-2)-x]^2\\
&-[z+pl(a-p)+t(2ap-p^2-1)-pm]^2\}.
\end{aligned}
$$

[146]

(请注意一个表面的悖论：该式看来可以分解成两个因子. 实际的情况是：只有当因子 $k+2$ 为素数而第二个因子等于 1 时，该式刚好产生正值.)

一个漂亮的肯定结果，我们最好是以它作为本章的结束吧！

阅 读 文 献

关于希尔伯特 1900 年在国际数学家大会上的著名报告和希尔伯特第十问题的解决情况，均可参阅 Felix Browder 所编：Mathematical Developments Arising from Hilbert Problems (American Mathematical Society, 1974)，丛书"Proceedings of Symposia in Pure Mathematics", Vol.28.

[147]

(李文林译)

第7章 四色问题

计算机数学时代来临

1976年,伊利诺大学的两位数学家,阿倍尔(Kenneth Appel)和哈肯(Wolfgang Haken),宣布他们解决了一个已经有整整一个世纪历史的、与地图着色有关的问题.他们说他们证明了四色猜想(four-colour conjecture).这本身就是一件值得报导的新闻.四色问题可能是仅次于费马最后定理(参见第8章)的最著名的未解决数学问题了.然而对数学家们来说,整个事情最有戏剧性的方面是在于证明该问题的方法.两位数学家的论证有很大部分并且是关键的部分是由计算机完成的,其中用到的一些概念本身就是计算机证明的产物.证明过程所需的计算量如此巨大,对它进行逐步检验已非数学家人力所及.这就意味着整个"数学证明"的概念发生了突变.自从50年代初电子计算机发展以来一直在酝酿着的事情终于发生了:计算机从数学家手中接过了从事真正的数学证明的重任.

在此之前,人们始终把证明看作是一种逻辑上严密可靠的推理过程,借助于这样的推理,数学家就可以使其他人相信某个判断的真实性.一个数学家在看懂一个证明以后,就会对所论命题的真实性深信不疑,同时也就会豁然理解确保这种真实性的推理.事实上,一个证明之所以成为证明,恰恰是因为它提供了这样一些推理. [148]

上述关于证明的简单观点,对于像单群分类定理(在第5章中作了介绍)这样极为冗长的证明可能需要作适当的引申,因为普通的数

学家在阅读一个写满两大卷五百页纸的证明时,通常会跳过许多的细节.但这实际上只不过是一种省劲的做法而已.既然相信别人已对证明的各部分作过检验,忙碌的数学家就没有必要再重复检查每一步细节.这样的证明仍然只是单纯的人脑劳动产品.虽然单群分类定理的证明有一部分也用到了计算机,但它们所得到的结果全都可以用人工来验证.计算机在这里并没有起"实质性"作用.

然而在四色猜想的证明中,计算机的使用却绝对起了实质性作用——整个证明直接依赖于计算机.为了接受这个证明,你必须相信所用的计算机程序确实做了其作者所说的事情.当阿倍尔和哈肯将他们的证明交给伊利诺数学杂志发表时,编者作了这样的安排:让人在另一台机器上用独立编制的计算机程序来检验证明的计算机部分!因此这个证明的关键部分对人们来说依然是藏而不见的.

许多数学家最初都发出了怀疑的声音.有人批评道:"整个过程主要是利用了计算机获得的结果,而这些结果从本质上来说又不能经受人工检验,这样的过程是不能被看作为数学证明的."对这些人来说,四色问题仍然悬而未决.事实上,能不能找到一个"标准"证明的问题至今没有解决.面对着繁杂不堪的计算,即使是计算机证明的支持者也不得不承认反对意见并非完全没有道理.迄止本书写作之时,也就是在首次宣布证明四色定理十多年后的今天,人们还不时听到这样的传闻,说是在计算机程序中已发现有一个重要错误,原先的证明将不能成立.但总的说来,随着时间的推移和计算机在社会中日趋增长的应用,拒绝接受四色定理计算机证明的数学家人数已在逐渐减少,大多数人现在认识到:计算机的出现不仅大大地改变了数学研究的方式,而且从根本上改变着证明概念本身.对产生"证明"的计算机程序的检查,现在应该被允许看作是一种有效的数学证明.

那么,这个对数学的本质有如此深刻影响的问题究竟是什么呢?[149] 故事得从头说起,那正好是第一台商用计算机问世前一百年的事情.

古色利的问题

1852 年 10 月的一天，刚从伦敦大学学院毕业不久的青年数学家弗兰西斯·古色利(Francis Guthrie，此人后来成为好望角南非大学数学教授)正在为一张英国地图着色.他当时发现，为了给任何一张 [150] 地图(平面图)着色并使其满足一个很自然的要求，即任意两个具有公共边界线的区域(国家、县城等等)着色不同，似乎最多只需要四种颜色就够了(见图 31).古色利不能证明这一事实，于是便写信把这问题告诉他的弟弟弗雷德里克(Frederick)，后者当时仍在大学学院学习物理.弗雷德里克又转而向他的数学老师，杰出的英国数学家德

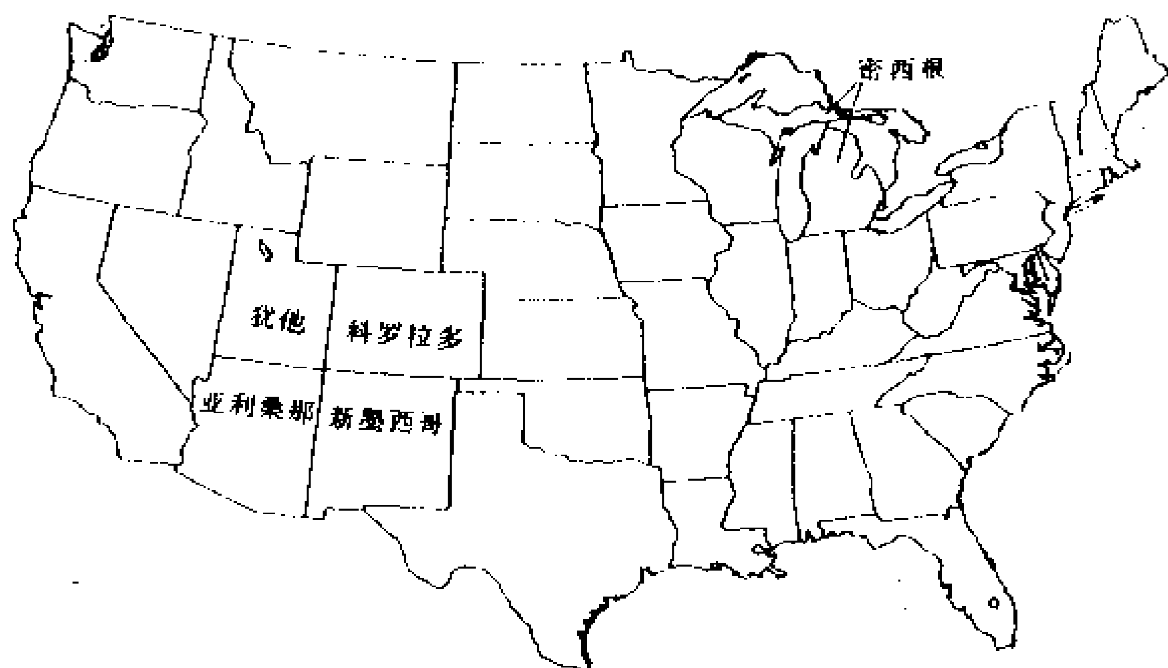


图 31 美国地图.用四种颜色就能给图中所有的州着色，并使得任何两个有公共边界的州着色不同.这样科罗拉多州与新墨西哥州(例如)必须着色不同，虽然科罗拉多州与亚利桑那州可以着色相同，因为它们只有一个公共点.在数学上，像密西根州这样由两个分离部分组成的州，必须被看成是两个不同的区域.你应该能毫无困难地证明：这张地图不可能只用三种颜色来着色(在上述特定的意义下).

·摩尔根(Augustus de Morgan)请教.

像弗兰西斯·古色利一样,德·摩尔根很容易就证明了至少需要四种颜色(即存在有这样一些地图,对它们着色只用三种颜色是不够的,见图32).他同时还证明了(参见下文):对于五个国家来说,不可能出现这样的情况,即它们中每一个都同时与其他四个相邻.这个事实乍看似乎蕴含了四色定理,但正如德·摩尔根本人大概已有所认识那样,实际上根本不是这么回事(见图33).(四色猜想从1852年首次提出,到1976年最终获证,其间出现的许多错误"证明",都是基于这一实际上并不成立的蕴含关系.事实上弗兰西斯·古色利本人有一个时期好像也陷入了这口陷阱.)

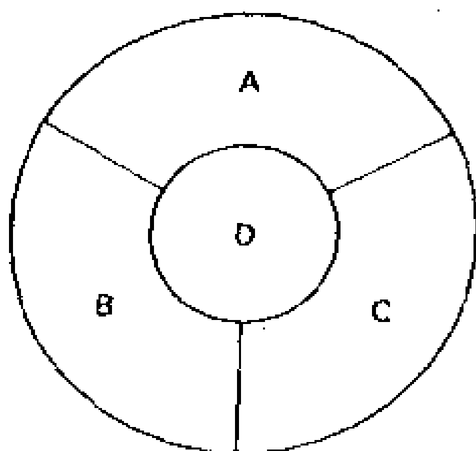


图 32 三种颜色不够.为了给所示的地图着色,使其中任意两个相邻国家着色不同,你必须用四种不同的颜色给A,B,C,D四国上色.

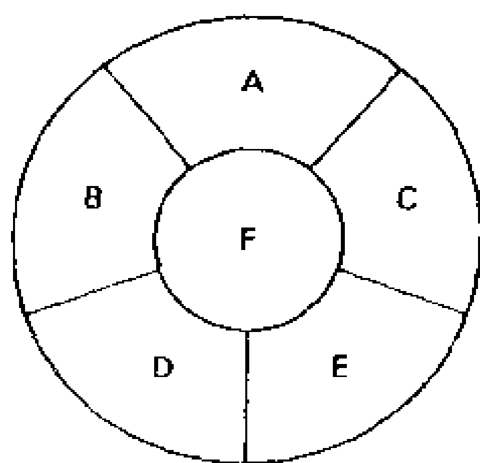


图 33 一个错误论证的示意图.许多人曾经以为:因为任何地图都不允许有这样的构形,其中五个国家每一个都与其他四个相邻,所以用四种颜色就足以给地图着色.这种蕴含关系实际上并不成立.在所示地图中,不存在这样的构形,其中四个国家每一个都与其他三个相邻,然而这张地图整个来说却不能只用三种颜色来着色,因此给一张地图着色所需的颜色种数与图中相互毗邻国家的最大个数并不恒等.

德·摩尔根未能解决这个问题，又把它转提给自己的学生和其他数学家（其中包括四元数的发明人，都柏林三一学院的哈密顿爵士（Sir William Hamilton——参见第3章），同时称赞了提出这个问题的古色利. 但总的来说这问题在当时并没有引起很大的兴趣，直到 [151] 1878年6月13日，英国数学家凯莱（Arthur Cayley）在伦敦数学会会员集会上当众发问是否有人能证明四色猜想.（凯莱的问题后来发表在伦敦数学会文集上，成为最早提到四色问题的印刷资料.）此举擂响了攻克四色问题的战鼓.

地图，网络与拓扑

任何人如果想要证明四色猜想，他所面临的第一个主要的困难就是问题涉及所有的地图——不仅是全世界所有地图集中的所有地图，而且包括所有可以想象的地图，上面有上百万个（甚至更多）形状、大小各异的国家. 只知道有某些特殊的地图可以用四种颜色着色对你并没有什么帮助. 你需要给出一种对所有情形都适用的证明. 这 [152] 意味着可以利用的条件很少. 那么究竟从何入手呢？眼下最明智的办法还是让我们先弄清楚问题包含着哪些可以肯定的东西.

为了讨论古色利问题，我们将“地图”定义成平面上任意多个区域（如果你愿意的话也可以叫“国家”）组成的图形，这些区域通过线（或称“边界”）而相互区分. 这个一般的定义包括了像图31那样的真实世界的地图，同时也包括了如图32，33和34所示的那种人造的“数学”地图. 关于图31所示的美国地图，实际上还存在一些潜在的问题，其中有些州占了两个不同的区域. 例如图中的密西根州就由被密西根湖分开的两个区域组成. 因为它们在物理上是不同的区域，当考虑地图着色问题时当然也应被认为是不同的区域. 同样地，长岛、纽约市也应该被看成是与纽约州其余部分不同的实体. 因此在地图的数学研究方面，占统治地位的是几何概念而不是政治观点.

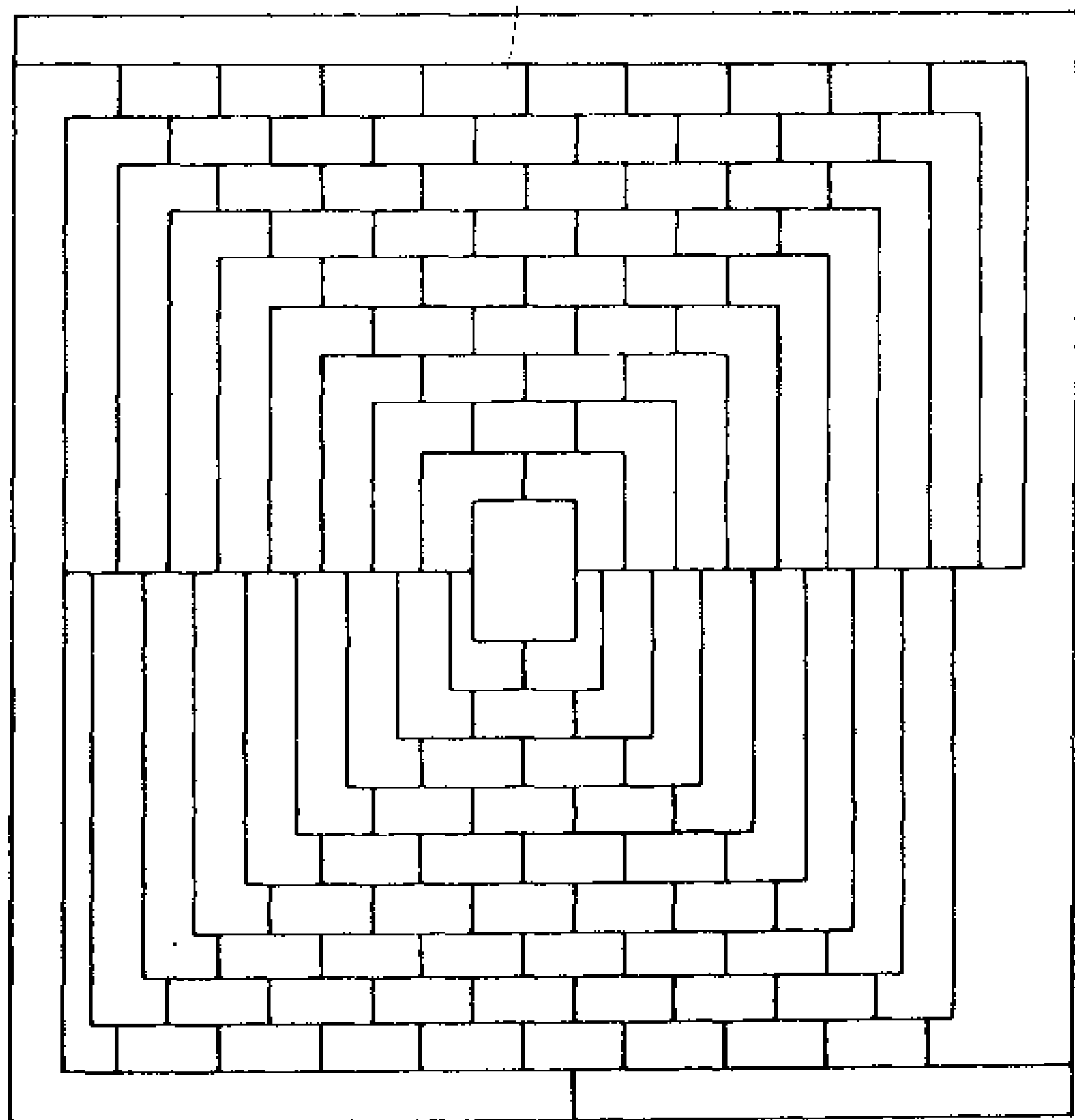


图 34 一张很难(虽然不是不可能!)只用四种颜色着色的复杂地图.(不妨试一试!)这张特别的地图是作为1975年4月1日《科学美国人》杂志愚人节玩笑的一部分而发表的.它也是著名的数学专栏作家马丁·加德纳(Martin Gardner)一篇文章的附图,加德纳煞有介事地宣布:这张地图是历史悠久的四色猜想的一个反例!

[153]

四色猜想讨论地图上区域(地理的而非政治的!)的着色,要求任意两个有公共边界的区域着色不同.(只有一个交点的区域,如图31中的亚利桑那州和科罗拉多州,不作为具有公共边界的区域,因此可

以着相同的颜色.)问题是求用这样的方式给地图上所有区域着色所需要的颜色的最少种数.这正是问题的第二个主要困难所在.即使是对一张特殊的地图,也会有许多种不同的着色方式,这里重要的不是任意特定的着色所需要的颜色数,而是使某种着色有可能的最少颜色种数.

稍加思考你就会认识到:对于地图着色问题来说,各个区域的实际形状与大小并不重要,重要的仅仅是它们的相对位置.这样,图35中所有的地图对地图着色员来说都是等价的.关于这一点,数学家的说法是:问题的实质在于地图的拓扑结构.

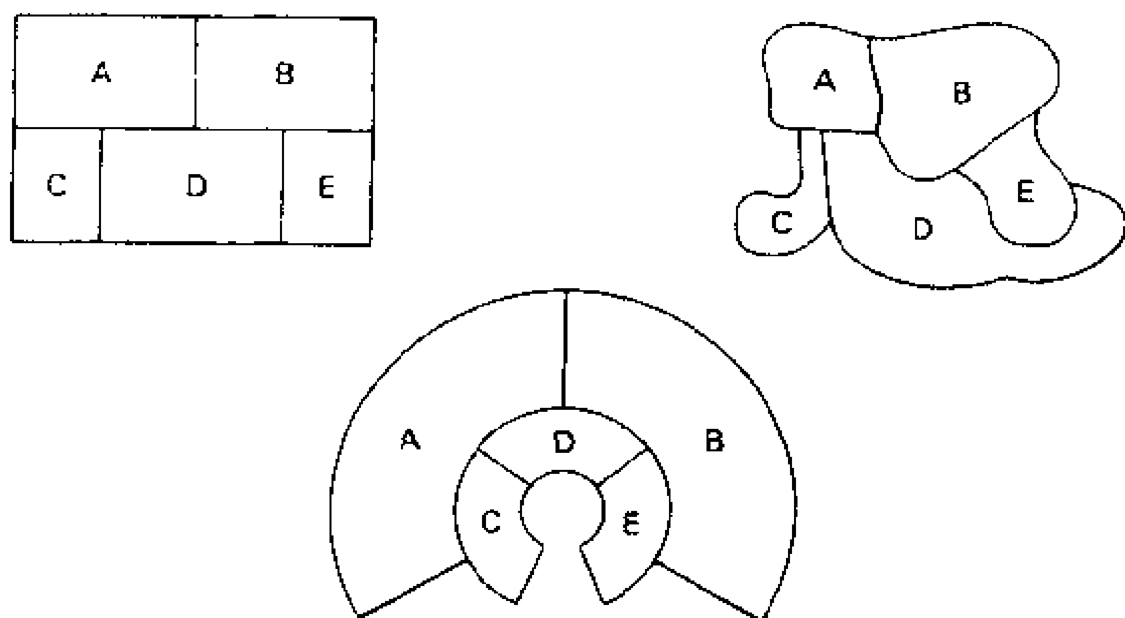


图35 拓扑等价性.就四色问题而言,图中所示的每张地图都互相等价;它们之间在拓扑上没有任何区别.

拓扑学是一门像几何学那样的数学分支.几何学研究二维、三维或更高维空间中的对象(或图形)的性质(当然,在四维或更高维空间中所谓"对象"具有高度抽象的意义),拓扑学也一样.二者的区别是在于所研究的性质的类型.在拓扑学中距离、大小是无关紧要的,直线、圆、角度等也失去了意义.事实上,拓扑学对普通几何中有根本意

义的性质全不考虑，而代之以研究对象(图形)在诸如弯曲、拉伸、压缩或扭转等连续变换下保持不变的性质．人们有时把二维拓扑学称[154]为"橡皮膜几何"，因为它研究的是图形的这样一些性质，如果你把图形画在一张"完全弹性"的橡皮膜上并且进行拉伸、扭曲等等，这些性质都不会改变(见图 36)．

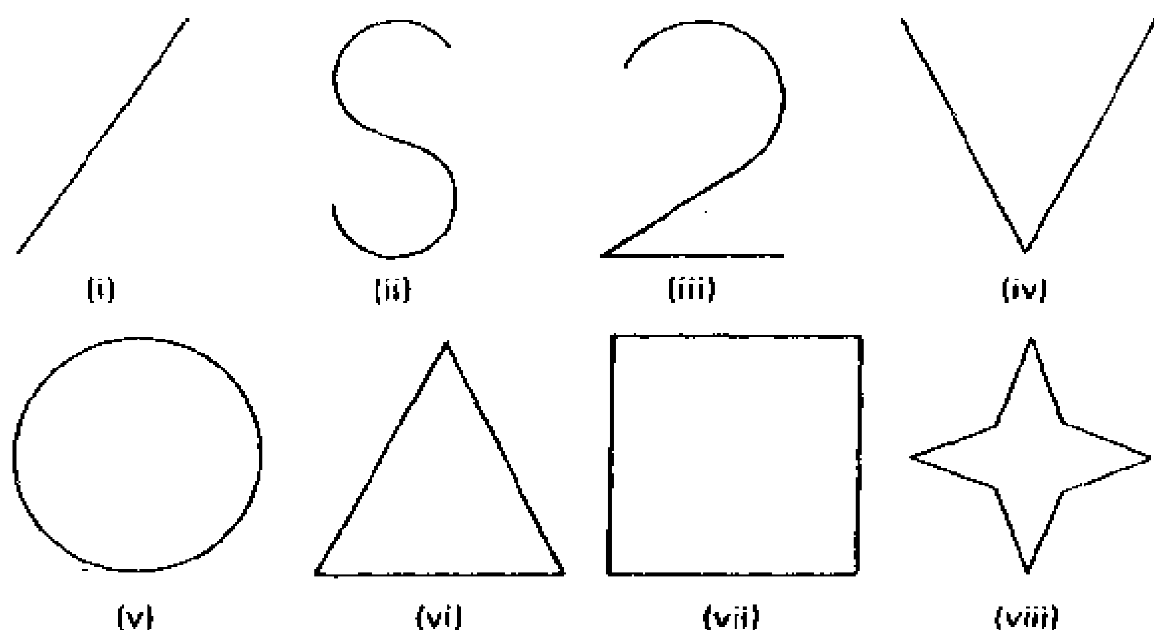


图 36 二维图形的拓扑性质．图形(i)～(iv)在拓扑上都是一样的，图形(v)～(viii)在拓扑上也是一样的，但(i)～(iv)中的任何一个与(v)～(viii)中的任何一个在拓扑上都不等价．

初次遇到拓扑概念的人也许会觉得在这样的领域里很难进行合理的数学研究．但事实相反，拓扑学是一个富有深刻结果的广阔的数学分支(参阅第 10 章)．实际上四色问题本身就是一个拓扑问题，虽然其解决并未用到很高深的拓扑技巧．图 35 可以说明这一点：其中组成一张地图的国家的形状与大小并不重要，重要的是它们的布局．你将会发现，如果你牢牢记住这一点，即重要的是一张地图的拓扑性质而不是其外表"形状"，对于理解下文的内容是很有帮助的．

一旦明白了这一点，一张地图的邻网络的概念似乎是考察四色

问题的切实可行的途径.已给一张地图,所谓邻网络可以这样来得到(见图37):在该地图的每一个区域内取一点,称之为网络的结点(node).(如果愿意的话,你可以设想这些点就是那些区域代表的国家的首都.)用一定的方式将这些结点连起来就形成一个网络(你可以通过类似方法用铁路网络将不同的城市连结起来).连结法则是:两个结点当且仅当它们各自所属的区域具有公共边界时,才将它们 [155]

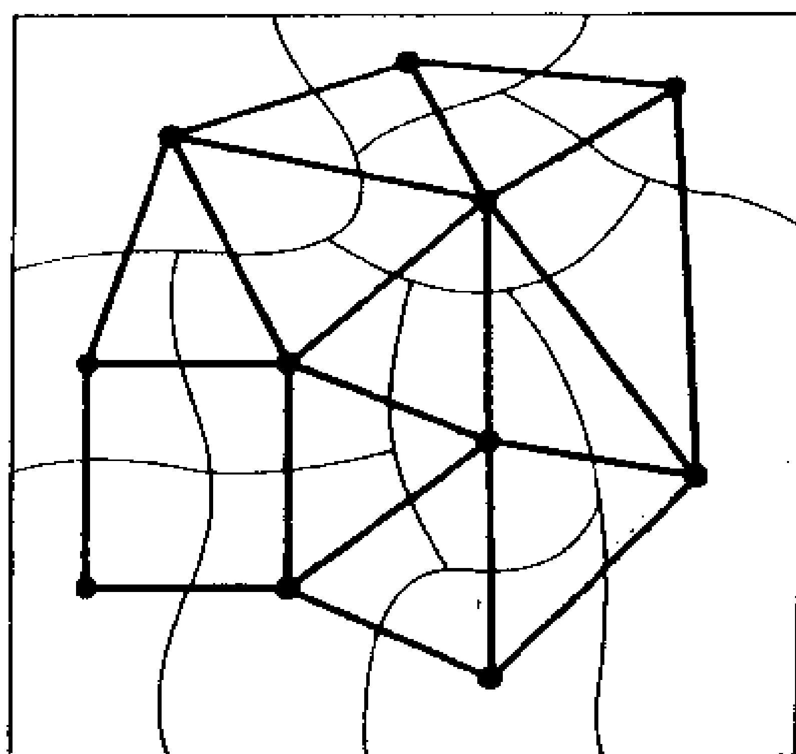


图37 邻网络.为了得到一张已知地图的邻网络,在该地图每个区域内取一点,用线将这些"结点"连结起来,而这些线则应完全位于相应区域的内部.这种连结只有当两个结点所属的区域具有公共边界时才进行,此时连线将跨过边界.因此连结就反映了公共边界的存在.将一张地图着色使任意两个相邻国家着色不同的问题,就等价于将相应的邻网络的结点着色,使任意两个有网络道路连结的结点着色不同.在本例中每对结点间都可能用直线相连结,但并不是在所有情况下都能做到这一点,曲线连结是允许的.(直与曲并不是拓扑性质.)

连结起来.这时连线必须完全位于这两个区域的内部并跨越公共边界.(对铁路网来说,这就是意味着铁路线不能穿过任何第三国的领土.)

邻网络清楚地显示了它所代表的地图的拓扑性质.确实,地图着色问题(在古色利问题的意义上)可用网络着色的语言重述如下:用这样的方法给网络的结点着色,使得任意两个相连的结点着色不同.如果所有的网络都可以用四种颜色着色,那么所有的地图也都如此,并且反之亦然.这样四色问题的网络表述提供了研究这问题的新的方法,它完全等价于原始的表述方法.于是研究这类网络就具有重要的意义.

这样就把问题带进了所谓图论(graph theory)的领域.注意根据邻网络定义的方式,网络中任意两条道路都不能交叉(或相交).一个[156]图(graph)与邻网络类似,不过取消了不允许道路交叉的限制.(此处"图"这个词的用法与它在数学上的另一个用法没有什么联系,后者涉及用"图纸"画曲线来表示方程.)虽然四色问题对图论这一数学分支的发展提供了许多原始的推动(从德·摩尔根那里知道这一问题的哈密顿就在图论方面做了许多早期工作),任意"图"的研究本身现在[157]已形成了一个广阔繁荣的数学分支.

欧拉公式

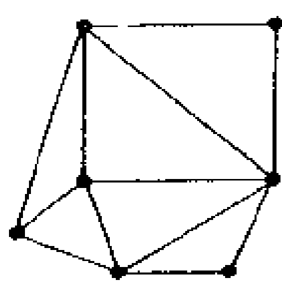
关于网络的一个特别有用的研究是由欧拉(Leonhard Euler)做的(背景略有不同).首先介绍一点术语(同时说明它们的来源).让我们一开始就约定,我们将只考虑具有这样性质的网络:从其中任何一个结点出发,都可能沿着一条纯粹由网络本身的道路构成的路线到达其他任何一个结点.这就排除了那些存在着没有连结道路的结点集的"病态""网络"的例子,却包括了对于研究地图着色问题来说所需的全部网络.任何一个这样的网络都将它所占据的部分平面划分成若干个区域,称这些区域为面(face).网络的结点有时(特别是当

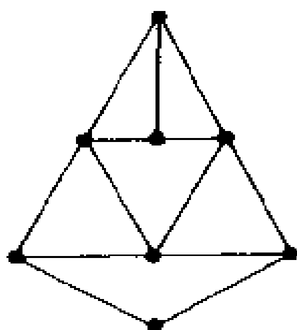
与欧拉公式有关时)叫作网络的顶点(vertex).连结这些顶点的道路叫边(edge).

现在你最好画出一些网络,对其中的每一个都标明其顶点数(V)、边数(E)和面数(F),如图 38 所示.然后对每种情形都计算出量 $V-E+F$,你将会发现结果总是等于 1.方程 [158]

$$V-E+F=1$$

对所有网络都成立,这一事实最先就是由欧拉本人证明的.





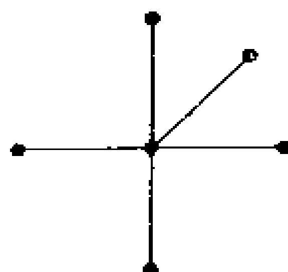


图 38 欧拉公式.对任何网络,顶点数(V),边数(E)和面数(F)都满足 $V-E+F=1$.

欧拉当初关心的主要是多面体而不是网络,这就解释了为什么会使用"顶点","边","面"这样一些名词.对任何一个多面体你将发现有

$$V-E+F=2.$$

(用多面体语言来说,这里 V,E 和 F 的意义是很明显的.)为了证明这个公式与我们刚刚得到的关于网络的结果实质相同,只需注意:如果你从一个多面体上挪掉一个面,然后将剩下的图形摊开在一个平面上,那么多面体原来的边就将形成一个连结原来的顶点的网络(这些顶点就是新网络的结点);反之如果有一张网络,你可以将它"撑"

成一个缺掉一面的多面体.当然正是这个缺少的面解释了网络公式[159]与多面体公式之间的差别.

欧拉网络公式的证明提供了一种在图论和四色问题的研究中都很有用的论证方法的出色例子.假定你从某个网络出发,希望证明 $V-E+F=1$.如果你从该网络挪掉一条外边(假设有这样的一条外边),那么会发生什么情况呢?这时 E 减少了 1,F 也是这样,而 V 则保持不变(见图 39).因此经过这样的操作后量 $V-E+F$ 保持不变.同样,如果该网络有一个"尾"(dangling)顶点(见图40),将这个点

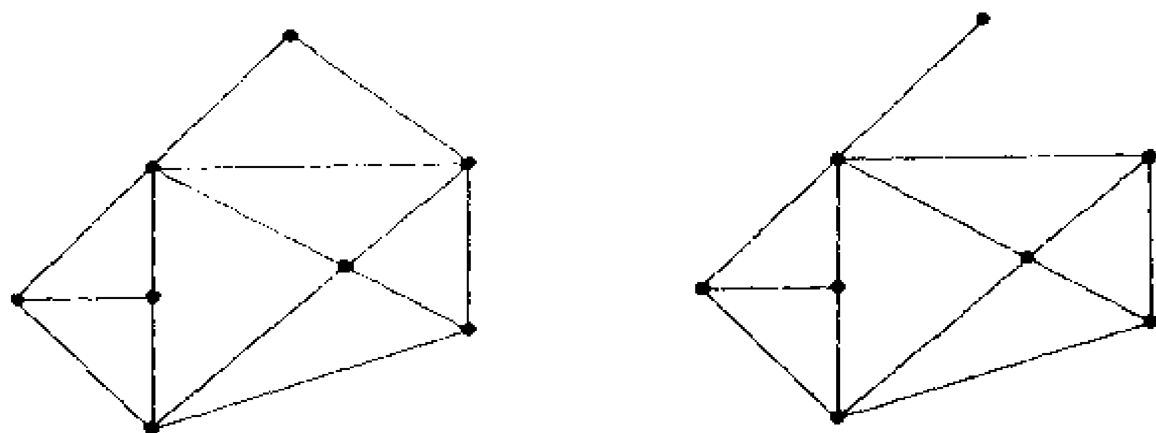


图 39 从一个网络挪去一条外边,E 和 F 都将减少 1,但 V 保持不变.这并不影响量 $V-E+F$ 的值.

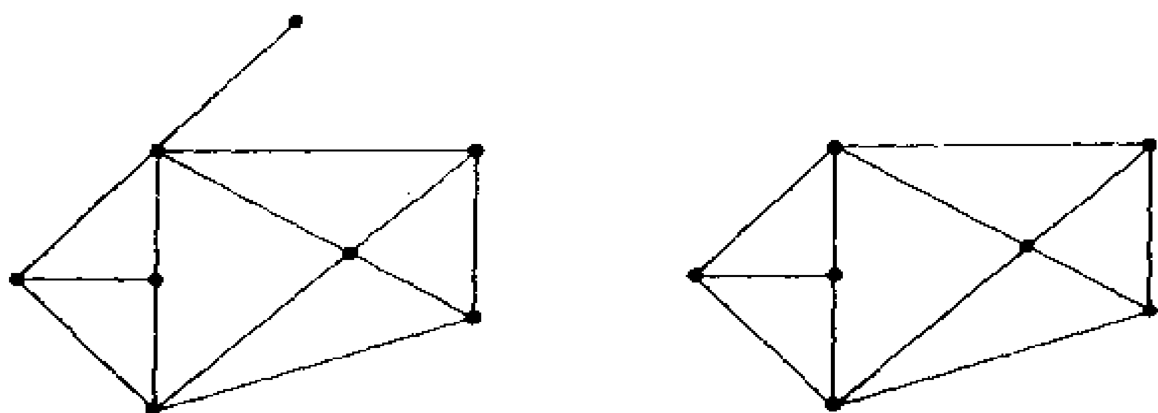


图 40 从一个网络中挪去一个"尾"顶点,V 和 E 都减少 1,而 F 不变.这样做并不影响量 $V-E+F$ 的值.

和通向它的边同时挪去，那又会发生什么情况呢？这时 V 减少 1，E 也是这样，F 则保持不变，在这一情形下量 $V-E+F$ 也将保持不变. 现在假如你从一个已知的网络出发，继续不断地挪去一切可能挪去的外边和尾点，就像大海吞蚀一个岛屿那样，最后你将得到一张只有一个顶点的网络. 也就是说你已经将最初的网络化约成平易(trivial)网络，其中 $V=1$，$E=0$，$F=0$. 在这个最后的网络中，量 $V-E+F$ 等于 1. 但是你所进行的"海蚀"过程始终不会改变量 $V-E+F$ 的值，因此在最初的网络中这个表达式的值也必定等于 1. 这就证明了我们的结论. 如果愿意的话，你也可以亲自来试一试. 从某个任意画出的网络开始，不断地挪去外边和尾点，同时逐步列出 V，E，F 和 $V-E+F$ 诸值. [160]

德·摩尔根定理

德·摩尔根关于古色利问题得到的一个肯定的结果是证明了：任何地图都不可能出现这样的情形，其中有五个国家每一个都与其他四个有公共边界. 用邻网络的概念和欧拉公式很容易证明这个结果. 用网络的语言，德·摩尔根的结果相当于说：不可能画出这样的网络，其中有五个顶点，每一个都与其他四个连结着. 当然如果你试图画出这样的网络，一定会发现最后总要留下两个顶点，你不可能将它们连结起来而又不穿过已经画出的某条道路. 但这并没有证明你的结论，因为之所以出现上述情况，很可能是已有的连结画得不合适. 下面给出的证明并不依赖特殊的作图，因此可以被看作是一个严格的证明. [161]

假设能够画出这样的一个网络，其中有五个顶点每一个都与其他四个相连结. 若将围绕整个网络的区域看作是一个附加的"面"，那么该网络的每条边都将划分出两个面. 并且因为现在多了一个面，所以欧拉公式变成

$$V-E+F=2.$$

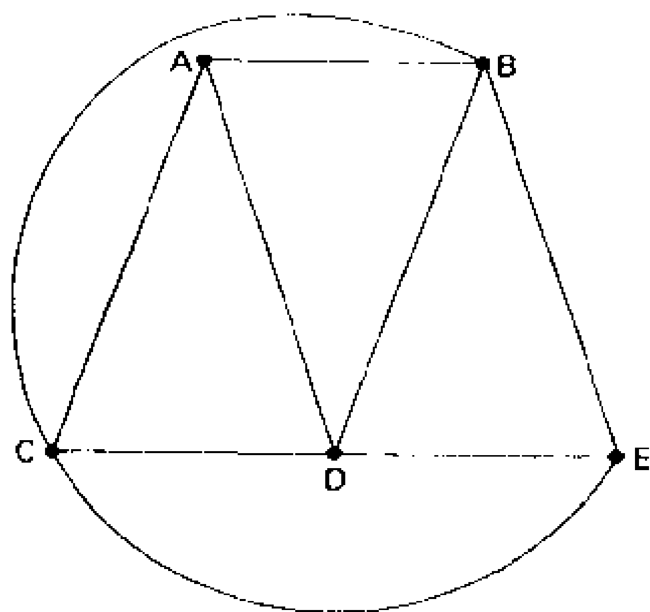


图 41 不可能画出这样的网络，其中有五个顶点每一个都与其他四个相连结．不管你怎样努力连结这些顶点，最后总会留下两个顶点（在本图中就是 A 和 E），你不可能将它们连结起来而又不穿过已经画出的某一条线．

我们已知这里的 V 值，它等于 5．同样，因为每个顶点都通过一条边与另一个顶点相连结，所以 $E=10$．（请你自己验算一下！）因此如果上述欧拉公式成立，则 F 必定等于 7．

到目前为止一切正常．现在我们来进行另一套运算．因为每个面都将至少被三条边所包围（这对我们刚刚引进的新面也成立，虽然在这一情形下“包围”这个词必须在拓扑意义下去理解），通过面来计算边，至少可以得到 $3\times7=21$ 条边．但是如果我们用这种方法（按面）来计算边数的话，每条边将被计算两次，因为它划分两个面．于是正确答案是至少有 $\frac{1}{2}\times21=10\frac{1}{2}$ 条边，这意味着至少有 11 条边（因为不可能有半条边）．但正如我们在前面指出的那样，$E=10$．这样我们就得出了矛盾，于是就像在这类证法中通常进行的那样可以得出结论说原先的假设是错的——这就是说，不可能画出这样的有五个顶点的网络，其中每个顶点都与其他四个相连结．这就证明了关于地图

的德·摩尔根定理.

五 色 定 理

1879年，在凯莱向伦敦数学会提出四色问题后不到一年，有一位叫肯泊(Alfred Bray Kempe)的律师，他也是伦敦数学会的会员，发表了一篇论文，其中宣布证明了四色猜想.但肯泊的证明是错误的，11年以后希伍德(Percy John Heawood)指出了其中一个严重的错误.不过经过希伍德补救以后，肯泊的方法可以用来证明有五种颜色肯定够了，这条“五色定理”的证明相当简单，可以在这里进行介绍. [162]

首先注意，通过上述将网络欧拉公式与多面体欧拉公式联系起来的推理，我们可以得出这样的结论：(就四色或五色定理而言)将地图画在平面上还是画在球面上是无关大局的.如果我们从一张画在球面上的地图开始，可以通过下面的做法将它变形成一张等价的平面地图：在其中某个区域的中央刺一个洞，然后将整个地图摊平(使得刺洞的区域包围了地图上的其他区域).反之，如果我们有一张平面地图，那么可以把包围原有地图的区域看成是一个附加的国家，然后把整个地图围成球形(将附加的围绕区域合拢起来形成一个像所有其他区域一样的“闭”区域).这个过程表明，如果每个平面地图都可用 N 种颜色着色，那么每个球面地图也都可用 N 种颜色着色，并且反之亦然.

实际上我们将对画在球面上的地图来证明“五色定理”，并将利用欧拉公式，对于球面地图来说这公式是

$$V - E + F = 2.$$

这里欧拉公式是被应用于地图本身，而不是像证明德·摩尔根定理时那样应用于相关的邻网络.(因此面是指地图的区域，边是边界，顶点则是三条或更多条边界线的交点.)

证明思路是这样的：给定一张画在一个球面上的(完全任意的)地图，用将两个或更多个相邻国家合并为一的过程逐渐地修改它，最

后得到一张至多有五个国家的地图——对它显然可以用五种或更少的颜色来着色.只要修改过程中所使用的每步操作不减少地图着色所需要的颜色数,这方法就能证明对最初的地图着色用五种颜色就够了.因此这种证明方法的关键是在于描述将已知地图约化为更简单的形式(即有较少国家)而又不减少所需的着色颜色数的特殊过程.共有六种不同的约简过程,每一种应用于由地图上特定的国家构形(configuration,即布局)决定的不同情形.

首先,如果一个区域完全被其他区域所包围(见图42(i)),那么可以将里面的区域与包围它的区域相合并.任何至少要用两种颜色的新地图的着色,都可以拓展成原地图的需要同样种数颜色的着色:只要给里面的区域着上不同于修改地图中合并区域的颜色就行了.

163

第二种约简被应用于这样的情况,即有一个顶点是至少四个区域的接触点.如果有四个或四个以上的区域在一点相接触,那么这些区域中必定有一对区域没有(在地图上任何地方)公共边界线(关于这一点可能需要你稍费思索),而这两个区域就可以合并为一个(见图42(ii)).给了修改地图的一种着色,原地图也就可用同样多种颜色来着色,只要使被合并的两个区域在原图中着相同的颜色,而其他区域的着色在两种情形都保持一样.反复应用这一约简过程,就可以将地图修改成在每一顶点都只有三个接触区域的情况.在以下的约简中,我们将假设已经达到了这种情况.

如果一个区域只有两个相邻区域(见图42(iii)),那么可以使它与其他两个区域中的一个相合并,而将它"取消".如果新的地图可以用至少三种颜色着色,那么原地图也可以用同样多种颜色来着色,只要给后被取消的中央区域着上与两个外围区域都不相同的颜色.

任何区域如果有三个相邻区域的话,都可以通过将它与邻域之一合并而被"取消"(见图42(iv)),并且如果新地图可以用至少四种颜色着色,那么原地图也必可用同样多种颜色着色,做法与上一种情形一样.

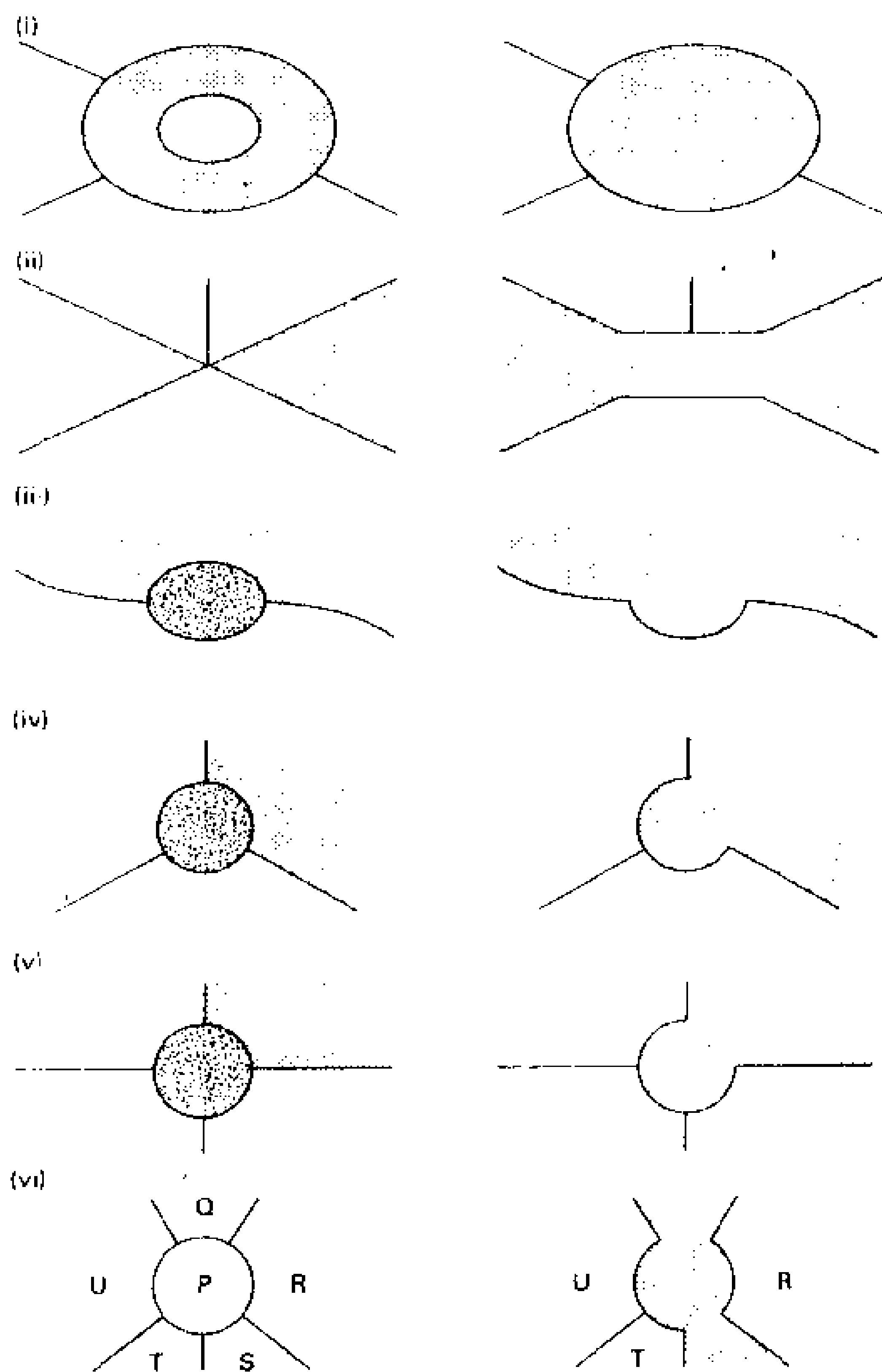


图 42 证明五色定理所使用的约简过程(详见正文).

同样，任何有四个相邻区域的区域可以与它的一个邻域合并（见图42(iv)），在可以使用五种颜色的情况下，这样做不会引起所需着色数的任何改变.

通过尽可能地使用上述各种约简过程，最后可以得到这样一张地图，其中任何区域都不会被其他区域完全包围，其中每个顶点都恰好是三条边的交点，并且其中每个区域都至少有五条边. 实际上，我们现在就要来证明：至少存在着一个区域其边数恰好为5.

设有 V 个顶点，E 条边，F 个面. 设 a 是界定每个区域的边数的平均数（因此 a 可能是分数）. 因为每条边都在两个区域之间，故

[164]

$$2E = aF.$$

同样，每个顶点都有三条边相交，每条边又连结着两个顶点，故

$$3V = 2E.$$

因此

$$3V = 2E = aF.$$

在欧拉公式 $V - E + F = 2$ 中代入 $V = \frac{1}{3}aF$ 和 $E = \frac{1}{2}aF$，就得到

$$\frac{1}{3}aF - \frac{1}{2}aF + F = 2,$$

于是

$$a = 6 - \frac{12}{F}.$$

因此 a 小于6. 因为每个区域的边数平均值小于6，那么一定有某些区域的边数小于6. 但所有区域至少应有5条边，因此一定存在某些区域恰好有五条边，这就是我们想要证明的.

现在来考虑这样一个有五条边的区域P，如图42(vi)所示，其邻域为Q，R，S，T，U. P有一对邻域相互没有公共边界，设为Q和S. 将P，Q，S合并起来. 如果新地图可以用五种颜色来着色，那么原地图也可以用五种颜色着色. 使被合并的区域Q和S在原图中着相同颜色，这样包围P的区域用了四种颜色，剩下一种颜色用于P.

整个约简过程到此结束. 每一步约简都使地图的区域数有所减少，反复约简后最终可以得到一张至多有五个区域的地图. 因为任何

这样的地图显然都可用五种颜色着色，那么原来的地图也就可以用五种颜色着色.事实上，通过上述约简过程的逐步反推，确实可以用一种完全机械化的方式用五种颜色来进行地图的着色.（不妨请读者从一张中等复杂程度的地图开始，亲自试一试！）

上述证明中用到的逐步约简地图的方法，最早由肯泊在其试图证明四色猜想的有缺陷的工作中提出，从那时以来，所有证明四色猜想的认真尝试（包括最后的成功的尝试）实际上都采用了这种约简地图的基本思想.由于问题的最终解决实质上是肯泊方法的推广（虽然是相当根本的推广），对他的论证再多加些笔墨是值得的. [166]

肯泊的方法

上述五色定理的证明，是希伍德对肯泊的错误论证进行补救而给出的.肯泊本人误认为可以解决四色定理的证明则比它要复杂得多.（图 42(v)和(vii)所示约简过程中使用的推理在只允许用四种颜色的情况显然不能成立.）肯泊的论证过程可以粗略地介绍如下.

首先假设存在一张这样的地图，需要用五种颜色才能给它着色（也就是说四种颜色肯定不够），然后证明由此可以推出矛盾.第一步是定义正则地图的概念.所谓正则地图是指这样的地图，其中不存在完全被其他国家包围的国家，其中任何一点至多是三个国家的接触点.从一张需用五种颜色着色的地图出发，运用上节所述的第一、二种约简过程，就可以得到一张需用五种颜色着色的正则地图.因为已假定了需用五种颜色着色的地图的存在性，所以需用五种颜色着色的正则地图也必存在.当然可能会有很多个这样的地图，它们包含的国家数不同.其中至少将有一个地图所包含的国家数最小（对这类地图而言），肯泊试图通过对这种需用五种颜色着色的最小正则地图的研究来推出矛盾.

使用最小正则地图的要点是：有些区域较少的（正则）地图可以用四种颜色着色，因此如果能够发现一种约简过程将地图上的国家

哪怕只减少一个，而又不改变五色的要求，那么你就立刻可以得出矛盾来，因为被约简的地图不可能同时发生两种情况，即既可用四种颜色着色，又不能用少于五种的颜色着色.

肯泊(正确地)证明了在任何一个正则地图中一定存在着至多只有五个邻国的国家，也就是说在图42(iii)，(iv)，(v)和(vi)所示的构形中，至少有一种会在该地图的某处出现，然后他(错误地)论证说，如果一个需用五种颜色着色的最小正则地图包含一个至多只有五个邻国的国家，那么就可以将它约简成有较少国家的正则地图，其着色仍需用五种颜色. 如前所述，这样就引出了一个矛盾. 肯泊的论证对[167]有两个，三个或四个邻国的国家来说是完全正确的. 两个和三个邻国的情形的证明在上述五色定理的证明中已经给出. 对四个邻国的情形则需要一个不同的、更巧妙的证明，需要检查该构形周围的那部分地图，并且可能需要改变某些周边国家的着色. 这里碰到的困难并不是不能克服，但还真要费一番心计. 而正如希伍德指出的那样，肯泊的错误是出在有五个邻国情形(见图42(vi)).

尽管如此，肯泊的证明已经涉及到两个关键的概念，它们在问题的最终解决中发挥了作用. 其中一个概念是所谓不可避免构形集(简称不可避免集，unavoidable set)——它们是这样一组地图构形，使得任何需用五种颜色着色的最小正则地图都至少包含它们中间的一个.(肯泊的不可避免集是由图42(iii)，(iv)，(v)，(vi)所示的那些构形组成的.)另一个概念是所谓可约性(reducibility)：如果某种特定的构形在需用五种颜色着色的最小正则地图中出现，就可以减少地图中的国家数而导致如下的矛盾，即需用五种颜色着色的正则地图包含的国家比最小正则地图还要少. 这样一来，只要你能够证明不可避免集中的每一个构形都是可约的，四色定理就得到了证明. 肯泊的证明之所以失败，恰恰在于他的可约性证明对于其不可避免集中的四个构形之一不能成立. 而阿倍尔－哈肯的证明之所以成功，又恰恰在于他们仔细地分析了肯泊受挫的最后一种情形，结果是需要找出不同的不可避免集. 阿倍尔与哈肯最后构造出来的不可避免集包含有

约1500种构形，了解到这一点，他们所面临的困难就可想而知了．我们将在适当时候说明这样的集合是如何发现的．但最好还是让我们先来介绍一下在肯泊的错误证明和阿倍尔－哈肯的最终解决之间这段时间里，四色问题究竟取得了哪些进展？

希伍德公式

前面已经介绍过希伍德的工作．在发现肯泊1890年证明中的错误以后，希伍德将他一生中其余的60年光阴全都花费在这个问题的研究上．虽然没有能最终证明四色定理，但是他却成功地在平面和球〔168〕面以外的曲面上解决了类似的地图着色问题．他的解答涉及到所谓曲面的欧拉示性数（Euler characteristic）．设有一个曲面，比如说球面，环面，或者甚至是双环面（见图43），在上面画一张覆盖整个曲面的地图，那么不管你怎样画这张地图，量 $V-E+F$ 的计算结果都是一样的——这与画在球面上的地图情形相同（那里答数是2）．因此这个量不依赖于地图（因为在这里任何地图都与其他地图一样），而是依赖于曲面（不同的曲面给出不同的答数），我们称它为曲面的欧拉示性数．它是曲面的一个拓扑不变量，这就是说不管曲面在拓扑上如何变化，这个量都保持不变．环面的欧拉示性数是0；双环面的欧拉示性数是－2．对克莱茵瓶（Klein bottle）来说，欧拉示性数是0，与环面一样．所谓克莱茵瓶是一种奇怪的单侧无边曲面，它在三维空间中是

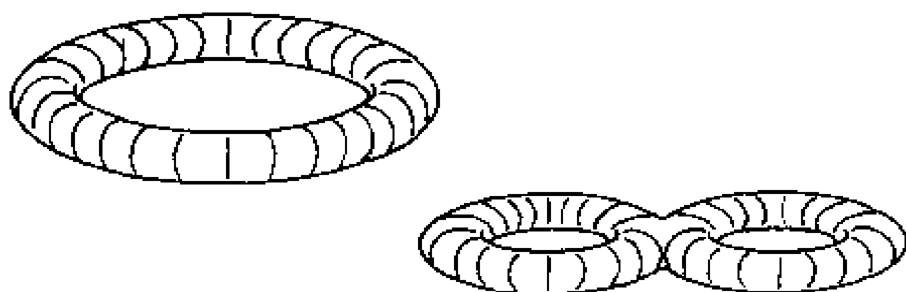


图43　环面和双环面

画不出来的，除非允许它自身相交(见图44). 但是请注意，虽然克莱茵瓶与环面的欧拉示性数相同，但它们在拓扑上却并不等价：你不可能将其中的一个变形为另一个. 拓扑等价的曲面欧拉示性数一定相等，但拓扑不等价的曲面其欧拉示性数则可能相等，也可能不等.

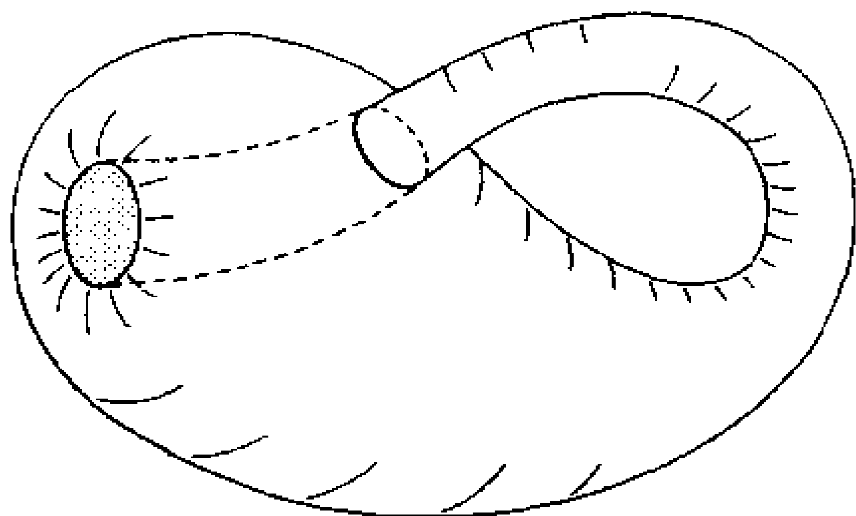


图44 克莱茵瓶——由无边单侧曲面构成的拓扑图形，在三维空间中只有在允许曲面自己"穿过"自己的情况下才能画出这个曲面图形来. 而在四维空间中这样的自相交就没有必要了.

利用与证明球面地图五色定理本质上相同的方法，希伍德证明了对于欧拉示性数为 n 的曲面，当 $n \leqslant 1$ 时，用

$$\frac{1}{2}(7+\sqrt{49-24n})$$

种颜色就可以给画在曲面上的所有地图着色. 不幸的是，$n>1$ 的唯[169]一曲面是球面，对球面有 $n=2$. 因此希伍德的漂亮结果恰恰在处理众所瞩目的情形时碰了壁——如果在希伍德公式中取 $n=2$，当然会得到答数4!

这样，对于环面($n=0$)来说，用七种颜色就可以给上面的地图着色. 同时，在环面上不难画出只用六种颜色不够着色的地图来. 因此我们就在较强的意义下证明了环面地图着色的"七色定理"：七种

颜色够了，再少了就不行．事实上，1968 年林格尔（Ringel）和杨斯（Youngs）证明了对于球面和克莱茵瓶以外的所有曲面，希伍德公式给出的是所需颜色种数的准确的最小值．（对于球面情形，当时还不知道最后的答案是什么．而对于克莱茵瓶来说，$n = 0$，希伍德公式给出的答数是 7，但有人另外证明了只需要六种颜色就够了[①]．）

向四色定理迈进

在希伍德之后，许多数学家（还有更多的业余爱好者）研究了四色问题，在这过程中发展起来的许多数学方法最终都在数学的其他领域中获得了应用．在所有这些努力中，有些对四色问题的最终解决确实可以被认为是起了引导作用，我们在这里作一简要的回顾．[170]

1913 年伯克霍夫（George Birkhoff）改进了肯泊的约简方法，并用以证明了比肯泊构形更大的一类构形是可约的．1922 年弗兰克林（Franklin）利用伯克霍夫的某些结果证明了每个有至多 25 个国家的地图都可以用四种颜色着色．1926 年雷诺德（Reynolds）将这一结果推广到 27 个国家，然后在 1938 年弗兰克林又创造了 31 个国家的纪录．1940 年温恩（Winn）证明了 35 个国家的情形以后，这方面的研究有所停滞，直到 1970 年，奥尔（Ore）和史坦普尔（Stemple）对所有至多包含 40 个国家的地图证明了四色定理．在阿倍尔和哈肯最终证明四色定理而使所有这类结果都黯然失色以前，这个数字曾经达到 96．

但是，尽管所有这些工作表明确有许多的构形是可约的，在 1970 年以前所有已被证明可约的构形的集合还远未能形成一个不可避免集（如四色猜想的证明所要求的那样）．人们构造了各种各样的不可避免集，但它们中似乎没有一个可以成为可约构形的不可避免集．你要么得到一个可约集，要么得到一个不可避免集，但却不能

① 因此，在四色定理被证明后，我们现在知道：克莱茵瓶是唯一的一种曲面，对它来说希伍德公式没有给出准确的最小颜色数．——原注．

二者兼得. 1950年,德国数学家希许(Heinrich Heesch),此人自1936年以来一直在研究四色问题,估计一个可约构形的不可避免集可能必须包含大约10 000个不同的构形. 虽然后来证明他的估计是过分夸大了,但它却正确地指明了:四色问题也许只有借助于能处理巨量数据的强有力的计算装置才能获得解决. 希许本人已认识到了处理庞大的构形集的能力可能是问题解决的关键,他事实上是第一个提倡并试图利用计算机来攻克四色问题的数学家.

希许从表述某些已知的证明构形可约的方法开始,他注意到了其中至少有一种(是肯泊方法的直接推广)可以完全机械化地在计算机上完成. 希许的一个学生杜勒(Karl Dürre)接着编了一个证明可约性的程序. 所有这一切都是利用表示地图的邻网络的语言来进行的,后者提供了在计算机上处理四色问题的更为方便的形式.

一个需要解决的问题是:关于一个特殊构形的可约性证明,一种方法失败了并不能说明这个构形一定不可约——在这种方法失败的地方也许另一种方法却能成功. 为了克服这一困难,需要发展一整套证明可约性的方法,也许可以称这套方法为可约性证明的"武器库". 早在1960年代后期,希许就已经为阿倍尔和哈肯1976年攻克四色问题的最后决战建造了一个相当巨大的武器库.

[171]

然而,在构造不可避免构形集方面也需要取得相应的进展. 希许曾提出一种方法,类似于移动电路中的电荷,但他并没有前进多远. 他本来应该继续前进的,因为这正是解决四色问题的法门!

希许的电荷法

与一个需用五种颜色着色的最小正则地图相应的邻网络是这样的网络(这可由肯泊的工作推得):它的每个面都是三角形,每个顶点至少是五条边的交点.(在一个顶点处相交的边数称为这个顶点的阶数.)希许的基本想法是把邻网络看成电路,并给每个顶点配置一个电荷. 如果一顶点的阶数为 k,给它所配的电荷量为 $6-k$. 这样阶数

为5的顶点就带有1个正电荷.6阶顶点不带电荷,7阶顶点带有电荷-1,如此等等.从肯泊的工作可以推知整个网络的带电量总和是12(对任意网络均成立).12这个具体数值并不重要,重要的是带电总量永远是正数.

现在假定你开始沿网络移动正电荷(如果愿意的话移动量可为分数).这当然不会引起网络带电总量的增减,但某些5阶顶点结果可能会失去所有的电荷(即被放电,discharged),而有些阶数大于6的顶点结果则可能带有正电(即被充电,charged).最后的确切状况显然将依赖于所使用的具体的重置(redistribution,或称放电discharging)过程.然而(这是关键之处),由于在不知道整个地图的情况下也可能确定其一小片局部的布局,因此给定一个特殊的放电过程(适用于任何地图)后,就有可能产生一张有限的构形表,它由所有最后只带正电的构形组成.

现在,因为网络带电总量为正,一定有某些顶点带有正电.这样,由于所有可能的正电荷的接受者都被包括在已由放电过程产生的有 [172] 限构形表中,所以(目前所考虑的)每个网络必定包含至少一个这样的构形.换句话说,所产生的构形表将形成一个不可避免集,这正是我们寻找的东西.肯泊当初的不可避免集可以被看作是完全没有电流移动的平易过程的产物.因此放电过程就是肯泊方法的推广.更进一步的推广应该能够带来更大的成功希望!

可以用一个简单的例子来说明上述内容(虽然要真正弄清是怎么回事,可能需要对这个例子进行反复的推敲和琢磨).假设放电过程是这样的:从每个五阶顶点向每个七阶或七阶以上的相邻顶点转移$\frac{1}{5}$单位的电荷,那么最后得到的不可避免集是如图45所示的只有两个构形组成的集合.为了说明为什么如此,首先请注意,一个五阶顶点只有当它至少有一个五阶邻点(图45(i))或六阶邻点(图45(ii))时,最后才会带正电.六阶顶点开始不带电,在这一过程中也不 [173] 接受电荷.七阶顶点只有当它至少有六个五阶邻点时,最后才会带正

电;如果是这种情形,则因每个面都是三角形,可以用一条边将这些邻点中的两个连结起来(于是图45(i)就适用于这样一对邻点).八阶或更高阶顶点即使当所有的邻点都是五阶时也不可能变成带正电的点.移动$\frac{1}{5}$单位的电荷是根本不够的.(例如一个八阶顶点如果有八个五阶邻点,起始电荷为-2,移入$8\times\frac{1}{5}=1\frac{3}{5}$单位的正电,那么最后带电量为$-\frac{2}{5}$.)

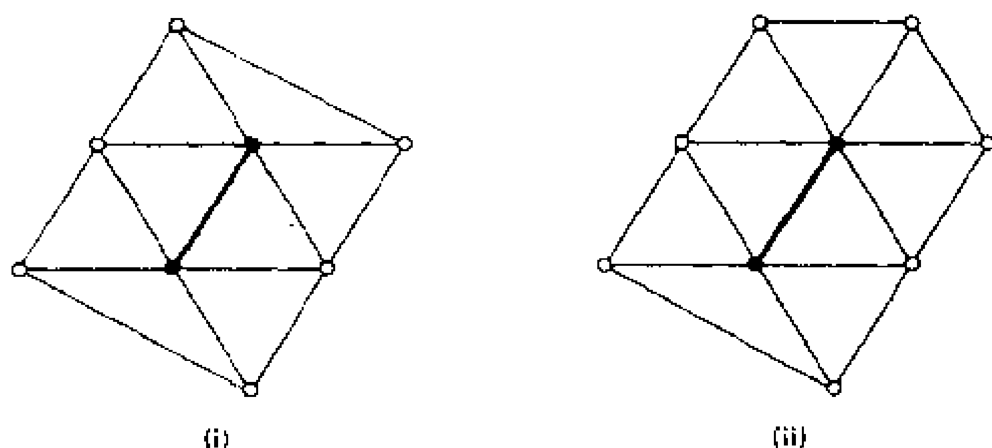


图45 由正文所描述的简单放电过程产生的一个不可避免集.该不可避免集由两个构形(i)和(ii)组成.(i)中的构形由两个连在一起的5阶顶点产生,(ii)中的构形则由一个5阶顶点和一个与它相连的6阶顶点产生.这两对顶点用黑圈表示,并用粗线连结.两个构形中其余的部分是根据产生它们的顶点对的阶数以及网络的面都是三角形这一事实来确定的.(对外侧各顶点的阶数没有限制,图中用白圈表示.)不可避免的意思是指任何地图中至少能找到两个网络(i)和(ii)中的一个.

这样图示的两个构形就形成一个不可避免集.就是说,因为上述论证对任意网络(对所考虑的类型)都成立,所以在任何地图中都可以找到这两个构形中至少一个.

利用电荷法证明四色猜想的基本想法是要找出一个放电过程,

使最后得到的不可避免集全部是由可约构形组成.如果能够做到这一点，四色定理将唾手可得.（顺便提一下，上述例子中所产生的两个构形都不是可约构形.）

四色定理的证明

那是在1970年，哈肯找到了一些新的方法来改进放电过程.面对着需要巨量计算时间的任务，似乎令人望而生畏.尽管如此，哈肯仍然希望这方面的研究最终将导致四色猜想的证明.1972年，哈肯开始与阿倍尔联手工作，全力以赴争取将希望变成现实.

他们的目标是要设计一种放电过程，通过它来产生一个由可约构形组成的不可避免集.这里涉及两件事：找出放电过程，同时证明它所产生的不可避免构形的可约性.他们首先研究限制较严的一类网络，这类网络用已有的成果（希许和其他人得到的）应该比较容易处理.总的思路是清楚的：从一个看似有望的放电过程出发，试图证明所产生的不可避免集中的每个构形都是可约的.如果表中的一个或某几个构形的可约性不能获得证明，就对放电过程加以修改，使那些有麻烦的构形不再出现.真是说起来容易做起来难！光是头两个放电过程的试算就花费了好几个星期的人机"对话".但事情还是在 [174] 逐渐地取得进展.正当两人苦心寻求改进的可约性证明方法之时，恰好有一种类似的通过人的介入来加强的计算机实验方法可供应用.三年之后，大约在1976年初，他们感到已经积累了足够的知识，可以发动最后的进攻了.作为全部实验工作的结果，他们发展了一种放电过程，可望能产生一个由可约构形组成的不可避免集.他们还编写了一个证明可约性的程序，似乎足以应付可能遇到的各种构形.他们使用的计算机程序具有自我修正的特性，一旦遇到一个不能证明其可约性的构形，正电荷就会自动地移动起来以求排除困难.但是这一切果真能行之有效吗？回答这问题的唯一办法就是到计算机上去通过程序，看看究竟会发生什么事情.他们是这样做了.

六个月以后，1976年6月，他们终于得到了回答．他们的程序(在两位非常内行的程序专家的帮助下)成功地证明了四色定理．前后经过整整四年的紧张工作，总共花费了1200个计算机小时．从最开始的放电过程到最后获得成功的放电过程，中间经过了500余次的修改，并且是不同的计算机运算的结果．两位数学家亲自用手算分析了10000多个正电荷顶点的相邻网络，而计算机则检查了2000多个构形并且证明了不可避免集中总共1482个构形的可约性．但一切都很顺利．一百多年的努力终于结出了硕果．

数学的面貌从此焕然一新！

阅读文献

关于四色问题获解的通俗易懂、非常全面的说明是 Kenneth Appel 和 Wolfgang Haken 为 Scientific American 杂志撰写的文章 the Solution of the four - color - map problem (Volume 237(October 1977), pp. 108 - 21).

为了回答认为他们的证明有缺陷的一些批评，这两位作者还撰写了另一篇关于他们的证明的文章(比上述那篇文章有更多的数学内容，但仍然是说明性的)，刊于 The Mathematical Intelligencer, Volume 8 (1986), pp. 10 - 20. 该文题为 The four color proof suffices(四色证明够了)，是由伊利诺地方邮局1976年为庆祝四色问题获解而使用的邮戳上一句话"four colors suffice"(四种颜色够了)改变而来，当时偶尔收到盖有上述邮戳的信件而事先又不了解情况的人肯定会感到莫名其妙．[175]

详细介绍四色问题及其解决情况的书有 Thomas L. Saaty 和 Paul C. Kainen 著的 The Four Colour Problem(McGraw - Hill, 1977).

N. Biggs, E. Lloyd 和 R. Wilson 所著 Graph Theory 1736 ~ 1936(Oxford University Press, 1976)一书提供了关于一般图论发展的良好的和全面的介绍．[176]

(李文林译)

第8章 费马最后定理

数学中最著名的定理

1983 年初,29 岁的西德数学家 G·法尔廷斯(Gerd Faltings)证明了一个结论,它标志着数学中最著名的未解决的问题取得了 100 多年来最重大的进步.这问题自然就是已有 300 年历史的难题——费马最后定理①;它的盛名远远超出了数学界,凡受过教育的人几乎没有不知道它的(虽然可能并不了解它的具体内容).这个问题的起源却只是潦草写在一本书的页边空白处的几句话.

皮埃尔·德·费马(Pierre de Fermat)于 1665 年 1 月 12 日去世时,是当时欧洲著名的数学家.尽管今天他的名字常跟数论为伴,可是由于他在这一领域的大部分工作超前了时代,致使他的同代人更多了解的是他有关坐标几何(费马独立于笛卡儿(Descartes)发明了坐标几何)、无穷小演算(牛顿和莱布尼茨(Leibniz)使之硕果累累)和概率论(这门学科本质上是费马和帕斯卡(Pascal)创立的)的研究.从总体上看,费马并非专业数学家,他的职业是律师兼土伦地方议会的推事②,1631 年他 30 岁时获得这个职位.

费马登上法学职位后开始业余的数学研究;虽然过去并未受过正规的数学训练,但他很快对数学产生了浓厚的兴趣.可惜费马没有 [177]

① 费马最后定理亦称费马大定理.——译者注.

② 推事:magistrate,负责审理案件的官员.——译者注.

养成发表成果的习惯，事实上在其整个数学生涯中，他未发表任何东西（除了极少数例外）. 另一方面，费马保持了跟同时代的最活跃和最伟大的数学家之间的广泛的通信联络. 在那个由数学巨人组成的世界里，有德沙格（Desargues）、笛卡儿、帕斯卡、沃里斯（Wallis）和雅克·贝努里（Jacques Bernoulli），而这位仅以数学为业余爱好的法国人能和他们中的任何一位相媲美：皮埃尔·德·费马堪称“业余者中的王子”.

著名的“最后定理”的生长道路既漫长又有趣. 1453 年，君士坦丁堡落入土耳其人之手，拜占庭的学者纷纷逃向西方，也带去了希腊学者的手稿，其中有丢番图的《算术》. 这部书一直留传至今，但在 1621 年前几乎无人去读它. 这一年，克洛德·巴舍（Claude Bachet）按照希腊原文出了新的版本，并附有拉丁文译文、注释和评论，这才使欧洲数学家注意到这本书. 似乎费马就是读了这本《算术》开始对数论产生兴趣的.

《算术》写于约公元 3 世纪，是丢番图的主要著作，而且是第一本见诸文字的代数书. 有关两个（或多个）变量的整系数方程的有理数解问题，是书中比较重要的部分. 今天的数学家在研究这些问题时，通常只限于求整数解，这对大多数情形没有本质上的差别. 如对于有 3 个变量的线性方程

$$2X+3Y+4Z=0,$$

其有理数解为 $X=\frac{1}{4}, Y=\frac{1}{10}, Z=-\frac{1}{5}$，若用 4，10 和 5 的最小公倍数 20 来乘它们，就得到整数解 $X=5, Y=2, Z=-4$. 同样的方法可用于其他许多情形，从而把有理数解转换成整数解. 对于本章所考虑的所有方程，上述转换都能实现，所以在大多数情形，我们将只考虑整数解.

在读巴舍版的《算术》时，费马喜欢在页边空白处写些简要的注记. 在他去世后 5 年，其子塞缪尔（Samuel）开始搜集父亲的笔记和信件准备出版. 他无意中发现了这本写有注记的《算术》，于是决定出版

此书的新版，并把费马写的页边注记作为书的附录. 48 条“对丢番图 [178] 的评注”(塞缪尔这样称呼那些注记)中的第 2 条，费马写在卷Ⅱ丢番图问题 8 旁的空白处，原问题是“给定一个平方数，将其写成其他两个平方数之和”，费马在旁写道(拉丁文)：

> 另一方面，不可能将一个立方数写成两个立方数之和，或者将一个四次幂数写成两个四次幂数之和. 一般地，对任何一个数，其幂次大于 2，就不可能写成同幂次的另两数之和. 对此命题我得到了一个真正奇妙的证明，可惜空白太小无法写下来.

用代数术语讲，丢番图问题想要求出满足方程

$$x^2 + y^2 = z^2$$

的 x，y 和 z，这是极容易的事. 费马在页边的注记断言，若 n 是大于 2 的自然数，则方程

$$x^n + y^n = z^n$$

不存在有理数解(如第 3 章所指出的，在丢番图时代，人们不认为 0 是个数，在某种程度上这也是费马时代的认识，所以平凡解——即至少有一个变量取 0 的解——不在考虑之列. 所以此问题仅涉及正有理数解).

注意，由上述简单的论证可知，对于丢番图问题和费马的断言，将有理数解转换成(正)整数解并无任何实质差别，因为任何有理数解都能导出一个整数解.(反之，任何整数解显然也是有理数解.)于是，我们可以将费马最后定理(回忆一下他的页边注记)说成是：对任意大于 2 的自然数 n，方程

$$x^n + y^n = z^n$$

无正整数解.

它为何叫“费马最后定理”呢? 这种称呼的来历多少有点含混. 我们不确切知道费马何时写下这一著名的页边注记，似乎是在他首 [179] 次研读丢番图的书的时期，即 17 世纪 30 年代的某个日子. 这是他的

数学生涯刚开始的时候，所以绝不会是他提出的最后一个定理．更大的可能是，他身后留下的众多定理中，这是最后一个留待证明的（如果它能被证明的话）！

上面解释了为什么用“最后”这个词．又为什么称其为“定理”呢？难道费马真的得到了他所说的“非常奇妙的证明”了？虽然存在这种可能性，但有证据表明他犯了错误，费马本人后来也认识到犯了错误．他提出的其他定理，或在信中，或在向其他数学家提出的挑战性问题中，都叙述或重复论述过．对 $x^3+y^3=z^3$ 和 $x^4+y^4=z^4$ 这两种特殊情形，他也在其他地方讲述过，唯独完整的最后定理只简单地出现在页边注记中．很可能他看出了如何去证 $n=4$ 的情形，可能也看出了 $n=3$ 时的证法（对这两个指数而言，定理肯定是对的），并认为他的论证方法可推广到其他任何整数 n 的情形，但其后发现情况并非如此．由于从未打算出版这些页边注记，费马可能觉得没有必要返回头来改动它．确实，他很可能完全忘了他写下了些什么！

尽管在普通人的心目中，相信费马真的找到了一个证明，但它毕竟只是个动人的故事：17 世纪的一位业余爱好者证明了一个结果，它使得其后（至少）350 年间的数学家为之奋斗、劳而无功．他的问题如此简明，因而这个故事更富感染力．而且永远存在费马是正确的可能性！

不论费马得到还是没得到证明，事实是尚无其他人能解决这个貌似简单的问题．这并不是缺少人问津造成的．许多大数学家在它身上绞尽了脑汁，对这个问题的研究已引出了全新的数学领域的发展（参见后面的内容），整本整本的书在讨论它（其中一些列在本章末尾）．确实，试图证明最后定理的努力得到了一系列意想不到的成果，费马最后定理对数学其他部分的意义已远远超出了定理本身．如果明天能证明费马最后定理，也就不再能从它导出新的数学成果了．它的重要性仅依赖于两件事：它的名声以及如下事实——一直没人能证明它！

[180] 本章将回答的问题是：我们现在对最后定理知道些什么？1983

年那个德国数学家取得的“最重大的进步”是什么？

毕达哥拉斯三元数

丢番图的《算术》中引出最后定理的那个问题是希望找出一种解方程 $x^2+y^2=z^2$ 的方法．由于该方程跟毕达哥拉斯(Pythagoras)定理有明显的联系，所以任何满足方程的三个整数 x,y,z 被称为一组毕达哥拉斯三元数．例如因为

$$3^2+4^2=5^2,$$

所以数 3，4 和 5 构成一组毕达哥拉斯三元数．一旦你得到一组毕达哥拉斯三元数，你就能从它导出无穷多组其他的毕达哥拉斯三元数，办法是用任取的数乘这三个数就行了．例如用 2 乘三元数 3，4，5 得 6，8，10，这也是一组毕达哥拉斯三元数，因为

$$6^2+8^2=10^2.$$

用 3 乘得 9，12，15 亦然．但从本质上看，我们此时只得到一组解，即 3，4，5，其余的皆是它的变种．另一方面，5，12，13 则是完全不同的另一组解(它照样也能导出它自己的解的家族)．3，4，5 和 5，12，13 这两组解跟通过乘常数导出的无穷多的解的差异在于，这些原始的解无公因子，即没有一个数能同时整除 3，4，5，也没有一个数能同时整除 5，12，13．

[181]

一般而论，若 a,b,c 是任一组毕达哥拉斯三元数，那么任一倍数 ma,mb,mc 亦然；反之，若 u,v,w 是任一组毕达哥拉斯三元数，又 d 是 u,v,w 的公因子，则 $u/d,v/d,w/d$ 也是毕达哥拉斯三元数．为了突出诸如 3，4，5 和 5，12，13 这样的“基本的”三元数的特性，数学家称无公因子(除了 1)的毕达哥拉斯三元数为本原毕达哥拉斯三元数．本质上，丢番图问题是要寻找一种能够得出所有本原毕达哥拉斯三元数的方法．

根据相当简单的数学推理，可以得到能生成所有可能的本原毕达哥拉斯三元数 x,y,z 的公式：

$$x = 2st, \quad y = s^2 - t^2, \quad z = s^2 + t^2,$$

其中 s, t 是任取的自然数，要求 s 大于 t. s 和 t 无公因子，s 和 t 中有一个是偶数，另一个为奇数. 例如取 $s = 2, t = 1$，三元数为 $x = 4, y = 3, z = 5$；取 $s = 3, t = 2$，则 $x = 12, y = 5, z = 13$；$s = 4, t = 1$，则 $x = 8, y = 15, z = 17$，等等.

欧几里得的《原本》(约公元前 350 ~ 300 年)中已有完整求解上述丢番图问题的内容.

$n = 4$ 的情形

假如已解决了丢番图问题，如何去证明费马最后定理本身呢？该定理(以整数形式表示)说：对任何自然数 $n > 2$，方程

$$x^n + y^n = z^n$$

无(正)整数解. 你如何证明(或试图去证明)这类命题呢？

明智的第一步是先研究几种特殊情形，如令 $n = 3, n = 4$ 或 $n = 5$；若能证明这些情形，你也许就会明白该如何去证整个定理. 看来，[182] 费马本人必定也是走的这条路，我们掌握的唯一具体的证据是他对 $n = 4$ 的情形所做的工作；实际上，这是人们所能见到的唯一的由费马本人所作的数学推理！它出现在《算术》的另一处页边注记中. 有趣的是，这个很特殊的注记又是这样结尾的："页边空白太小，我无法写出完全的证明和所有细节."

在解释费马的论证(以及它如何去证明 $n = 4$ 的情形)之前，你可能会自问如何(用一般的术语)对付(诸如)$n = 4$ 时的问题. 你可能会试几组 x, y, z 的具体数值，检查它们之中是否有满足如下方程的：

$$x^4 + y^4 = z^4.$$

(可以想象，你不希望找到解，因为费马就是那样说的.)在试过多组数值后(没有找到解)，你可能会去编一个计算机程序，以便对方程的解作更广泛和系统的研究. 比如对 x, y, z 试取从 1 到 100 的所有整

数(实际还有更有效的方法).但在计算机上算了几个小时,你仍不会成功;这种办法之所以无效是很显然的,因为无论你的计算机功能如何强大,也无论你的算法的效率有多高,这种战略在证明费马的论断(此时是 $n=4$)方面绝不会成功.最后定理($n=4$)断言,没有一组三元数是 $x^4+y^4=z^4$ 的解,为此你需要检验无穷多组三元数,而根本不存在一种计算机能使你检验无穷多种情况.这种方法倒可能用于否证最后定理,因为此时只消找出满足费马方程的一组解就够了.为了证明最后定理,或证明它的任何一种特殊情形,我们需要更精巧的数学方法.

数学中常用的一种最可靠的办法是通过导出矛盾来证明命题.当你想证明($n=4$ 时)不存在满足方程 $x^4+y^4=z^4$ 的解,此时可先假定它有一个解,比如 x,y,z,然后基于这个假设(用数学推理)演绎出矛盾来.一旦导出矛盾,你就达到了目的,因为矛盾只能来自错误的假设——你在这里的假设是方程存在一个解. [183]

我们面临的问题是如何从你的假设演绎出矛盾.对于包含自然数的任何陈述(如最后定理)而言,费马创造了一种特别有效的证明方法,即所谓的无限递降法.他说这是他在数论中进行所有证明的基础.费马在《算术》的页边空白处潦草书就的那个(上面提到过)注记,就是这种方法的一个说明,其中涉及著名的毕达哥拉斯三角形(你将立即看到它跟 $n=4$ 时的最后定理的联系).

据众所周知的理由,一个三角形被称为是毕达哥拉斯的(简称毕氏的),若它含有一个直角且三条边的长度皆是整数.(换言之,毕氏三角形是其边长构成一组毕达哥拉斯三元数的三角形.)费马证明的是:这样的三角形的面积绝不是平方数(即绝非整数的平方).他的证明如下.

假设存在一个毕氏三角形,其面积恰是某整数的平方.令 x,y,z 是三角形的边长,z 是斜边(见图 46),于是由毕达哥拉斯定理可知,x,y,z 满足恒等式

$$x^2+y^2=z^2.$$ [184]

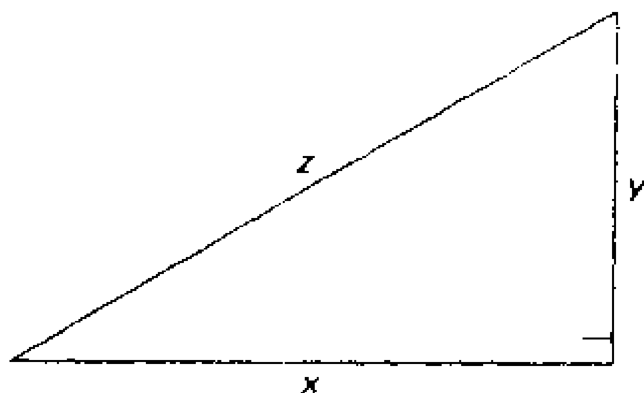


图 46 费马有关毕达哥拉斯三角形的结论.据毕达哥拉斯定理,直角三角形的边长满足方程

$$x^2+y^2=z^2.$$

此时,三角形的面积公式为

$$\text{面积}=\frac{1}{2}xy.$$

费马利用无限下降法,从假设 x,y,z 皆为整数,三角形面积是某个整数的平方出发,演绎出自相矛盾的结论.

令三角形面积为 u^2(u 是整数).根据面积公式,我们有

$$u^2=\frac{1}{2}xy.$$

依靠一种相当巧妙的论证,费马能导出另一组正整数 X,Y,Z 和 U,使得

$$X^2+Y^2=Z^2, U^2=\frac{1}{2}XY, Z<z.$$

(论证的细节见 H·M·爱德华兹(Edwards)关于费马最后定理的书的第1章,参见本章末的书目.)由此我们能容易地推导出所需的矛盾. X,Y,Z 具有 x,y,z 的所有性质,故可施以同样的论证而导出另外4个正整数 X_1,Y_1,Z_1,U_1,使得

$$X_1^2+Y_1^2=Z_1^2, U_1^2=\frac{1}{2}X_1Y_1, Z_1<Z.$$

类似地,又必存在另4个正整数 X_2,Y_2,Z_2,U_2,使得

$$X_2^2+Y_2^2=Z_2^2, U_2^2=\frac{1}{2}X_2Y_2, Z_2<Z_1.$$

这种推理过程可无限继续下去.我们称这种过程为无限递降,因为正整数 $z, Z, Z_1, Z_2, \cdots$ 越来越小(即 $z>Z>Z_1>Z_2>\cdots$).但这将引出矛盾,因为不存在无限下降的正整数序列:当你降到1之后就不得不停止这一过程.所以,结论只可能是不存在面积为某个整数的平方的毕氏三角形.

费马似乎是为了证明 $n=4$ 时的最后定理而构造了上述方法,尽管没有直接的证据表明他具体地找到了两者的联系.为了从上述涉及毕氏三角形的结论推导出 $n=4$ 情形下的最后定理,所需的仅是下述简单而巧妙的技巧.

设方程 $x^4+y^4=z^4$ 有一组解.根据解的值,我们取 $a=y^4, b=2x^2z^2, c=z^4+x^4, d=y^2xz$.反复利用熟知的代数恒等式

$$(r+s)^2=r^2+2rs+s^2,$$

[185]

你就得到(请自己检查一遍!)

$$\begin{aligned}a^2+b^2 &= (z^4-x^4)^2+4x^4z^4\\ &= z^8-2x^4z^4+x^8+4x^4z^4\\ &= (z^4+x^4)^2\\ &= c^2,\end{aligned}$$

并且有

$$\frac{1}{2}ab=\frac{1}{2}y^4 2x^2z^2=(y^2xz)^2=d^2.$$

于是,$a^2+b^2=c^2$,且$\frac{1}{2}ab=d^2$.但上面已经证明这是不可能的.因此所作的假定"方程 $x^4+y^4=z^4$ 有解"必定是错的,证明完毕.

由此立即可得,最后定理对于 n 为4的任一倍数时成立.理由很简单,若方程

$$x^{4k}+y^{4k}=z^{4k}$$

有解 $x=a, y=b, z=c$,那么 a^k, b^k, c^k 将是方程 $x^4+y^4=z^4$ 的解,而刚刚证明这是不可能的.

更一般地,若能对任一给定的指数 m 证明最后定理成立,那么该定理对指数为 m 的任一倍数时亦真.又因为所有大于2的整数或是可被大于2的素数整除,或是可被4整除(或同时被两者整除),所以为了完全证明最后定理,只需考虑 n 为大于2的素数(即奇素数)或 $n=4$.由上所述知 $n=4$ 的情形已经解决,问题便归结为 n 是奇素数的情形了.

进一步的研究往往将奇素数分成两种从属的情形.首先要注意,跟 $n=2$ 时(毕达哥拉斯三元数)的情形类似,若 x,y,z 是方程

$$x^n + y^n = z^n$$

的解,那么 x,y,z 的任一倍数也是解,所以真正要讨论的问题是:对于给定的 n 是否存在本原解,即无公因子的解 x,y,z.对于给定的 [186] 奇素数 p,(指数为 p 时的)最后定理的第一种从属情形是指:方程

$$x^p + y^p = z^p$$

不存在本原解 x,y,z,使得 p 不能整除 x,y 和 z 中的任何一个.第二种从属情形是指:方程不存在本原解 x,y,z,使得 p 能整除 x,y 和 z 中的一个.显然,对给定的 p,要证明指数为 p 的最后定理等价于证明跟这个 p 相应的两种从属情形.将一个命题分解为两种从属情形来证,其好处在于能在"分而治之"的基础上取得某些进展.我们将在本章后面部分再讨论这个问题.

如前所述,我们尚无证据说明费马确实对 $n=4$ 的情形证明了最后定理.上面简略提到的他对毕达哥拉斯三角形所证得的结论表明,他完全有能力据此证明 $n=4$ 的情形.无论如何,大多数资料使人乐于相信他获得了此项成果.有关 $n=3$ 时的最后定理的证明,同样被一片阴云所笼罩;尽管几乎所有的人都把它归功于L·欧拉,但无法得到确切的定论,因为在他有关这一证明的唯一出版物中同样存在漏洞.

$n=3$ 的情形

在 1753 年 8 月 4 日寄给 C·哥德巴赫(Goldbach)的信中，欧拉宣布他成功地证明了 $n=3$ 时的费马最后定理，但没有给出证明. 直到 1770 年，他才在圣·彼得堡出版的他的《代数学导论》中给出了一个证明. 他是否在 1753 年已经得到了证明，现不得而知，但 1770 年的证明肯定有严重缺陷. 不过，对于 $n=3$ 的情形，这一缺陷尚可补救；而对其他情形，类似的错误却无法克服. 欧拉的证明太长，无法在此给出细节，但为解释欧拉所犯错误的性质，有必要泛泛地描述一番，进而说明对其他情形的最后定理，这种错误是致命的.

如同费马对 $n=4$ 情形的"证明"一样，欧拉也使用了无限递降法. 他先假定 $x^3+y^3=z^3$ 存在解 x, y, z；然后推导出必存在另一组解 X, Y, Z，使得 $Z<z$. 整个论证的症结在于证明如下命题：若 p 和 q 是两个无公因子的数，又若 p^2+3q^2 是立方数，则必存在数 a 和 b，使得 $p=a^3-9ab^2$ 和 $q=3a^2b-3b^2$. 此命题完全正确，可用欧拉在别处提出的方法证明. 但在他发表的最后定理的证明中，欧拉采用了一种新型的论证法，其中涉及形如 $a+b\sqrt{-3}$ 的数(a, b 是整数)，这就招致了错误. [187]

我们能够理解欧拉为什么要用 $a+b\sqrt{-3}$ 这样的数. 先将 $(a+b\sqrt{-3})^3$ 展开，它等于

$$a^3+3a^2b\sqrt{-3}-9ab^2-3b^3\sqrt{-3},$$

作适当合并后变为

$$(a^3-9ab^2)+(3a^2b-3b^3)\sqrt{-3}.$$

所以，若 $p=a^3-9ab^2, q=3a^2b-3b^3$(正是被他证明的命题中提到的式子)，那么

$$p+q\sqrt{-3}=(a+b\sqrt{-3})^3.$$

注意，命题中的一个假设为 p^2+3q^2 是立方数，这等价于 $(p+$

$q\sqrt{-3})(p-q\sqrt{-3})$是立方数，故整个命题可重述为：若 p 和 q 无公因子，又若$(p+q\sqrt{-3})(p-q\sqrt{-3})$是立方数(在含有$\sqrt{-3}$的数系中)，那么 $p+q\sqrt{-3}$必为立方数(这又是在 $p+q\sqrt{-3}=(a+b\sqrt{-3})^3$ 的意义下的结论，此处 a,b 为整数).

为了证明重述后的命题，欧拉的推理如下：数 $a+b\sqrt{-3}$(a,b 是可以变化的整数)形成一个数系，它非常像由整数构成的数系(参见第3章). 若 m 和 n 是两个给定的无公因子的整数，且若 mn 是立方数，则 m 和 n 皆为立方数. 通过类比，欧拉认为这结论对由$\sqrt{-3}$构成的数系同样适用；因为如欧拉正确证明了的，若 p,q 无公因子，便蕴含了$\sqrt{-3}$数系中的数 $p+q\sqrt{-3}$和 $p-q\sqrt{-3}$亦无公因子 [188] (在$\sqrt{-3}$数系内)，于是立即可导出所希望的结论.

上述论证中的缺陷——一个严重的缺陷——是用跟整数的类比进行推理造成的，实际上这种类比根本不可靠. 虽然 $a+b\sqrt{-3}$这个数系在许多方面跟整数系相似(两个系统都构成整数环，参见第3章)，但它并不具备整数的全部性质. 欧拉的证明中用到的最要紧的性质，即唯一因子分解定理，对整数系是成立的(算术基本定理)，即每个整数可唯一分解成素数的乘积. 若你读过第3章，你会知道由$\sqrt{-3}$形成的数系也确实具备这种性质，所以欧拉的结论是正确的. 但是你从第3章也知道，总共只有9个整数根能保证唯一因子分解的性质，$\sqrt{-3}$恰是其中之一. 欧拉完全是靠运气，才使他用这种类比论证时没有导致错误. 如果他试图去证 $n=5$ 时的最后定理，就要利用 $a+b\sqrt{-5}$这样的数，那么他的方法就失效了. 我们即将看到，唯一因子分解的失效像一块暗礁，许多意欲获得的证明都会因它而失败.

再看两种情形：$n=5$ 和 $n=7$

1825年，P·狄利克雷(Peter Gustav Lejeune Dirichlet，他刚到20

岁)和A·-M·勒让德(Adrien - Marie Legendre,他已年过70岁)证明了 $n=5$ 时的最后定理.他们的方法基本上是欧拉对 $n=3$ 时所使用方法的延伸.跟在那里起关键作用的等式

$$p+q\sqrt{-3}=(a+b\sqrt{-3})^3$$

类似的等式是

$$p+q\sqrt{5}=(a+b\sqrt{5})^5.$$

然而,为了证明 $p+q\sqrt{5}$ 是五次幂数(此时肯定不能使用跟整数作类比来推理),他们不仅必须假定 p^2-5q^2 是五次幂,p 和 q 无公因子(跟 $n=3$ 时一样),而且还要假定 p 和 q 中只有一个是偶数,且 q 可被5整除.他们也没有使用唯一因子分解定理(因为此时它不成立). [189]

随着 $n=5$ 情形的解决,到那时为止大家都使用的方法,开始显现出不堪重负的迹象,对代数工具的要求变得越来越苛刻.1832年,狄利克雷在寻找 $n=7$ 的证明方法失败后,成功地证明了 $n=14$(这是个相当弱的结论)时最后定理成立.1839年,G·拉梅(Gabriel Lamé)终于证明了 $n=7$ 的情形,他必须诉诸于某些跟7本身结合得十分紧密的精巧工具;因此若不采用全新的手法,人们无望将他的方法应用于下一个 $n=11$ 的情形.正是拉梅本人于1847年提出采用新的行动路线的建议.

分圆整数和拉梅的宣告

拉梅的建议的核心是试图利用 n 次复单位根来彻底解决最后定理的证明.所谓 n 次复单位根是指一个复数 r,它满足 $r^n=1$,但 $r^k\neq 1$,k 为小于 n 的任意正整数(这里的 n 为任意一个奇素数).对任意取定的奇素数 n(实际上是对任意的奇数 n),数1有 $n-1$ 个 n 次复根.如对于 $n=3$,1的两个3次复根为

$$-\frac{1}{2}+\frac{\sqrt{3}}{2}i,\ -\frac{1}{2}-\frac{\sqrt{3}}{2}i.$$

(你可以通过计算出这两个复数的立方来验证.)

引进 r 的目的何在？到那时已找到的对 $n=3,4,5,7$ 等情形的证明，无一不要依靠代数中的某种因子分解，如对 $n=3$ 的情形利用了因子分解式

$$x^3+y^3=(x+y)(x^2-xy+y^2).$$

拉梅认识到，n 增大时证明的困难程度也增加，其原因在于进行这类因子分解时，被分解后的因子中有一个的次数越来越高. 引进 r 就有可能彻底地将 x^n+y^n 分解成 n 个因子，它们都是 1 次的.

[190]

为了进行因子分解，应注意到复数 $1,r,r^2,\cdots,r^{n-1}$ 是复方程

$$z^n-1=0$$

的根，所以有

$$z^n-1=(z-1)(z-r)(z-r^2)\cdots(z-r^{n-1}).$$

假定令 $z=-x/y$，并用 y^n 乘方程的两边，你将得到(因为 n 是奇数)

$$x^n+y^n=(x+y)(x+yr)(x+yr^2)\cdots(x+yr^{n-1}).$$

上式中 x^n+y^n 的每个复因子皆是下述形式的数的特殊情形：

$$a_0+a_1r+a_2r^2+\cdots+a_{n-1}r^{n-1},$$

其中 $a_0,a_1,\cdots,a_{n-1}$ 是整数. 这种类型的数——由整数和 r 的幂组成——现在称为分圆整数(所有这样的数仍然跟某个取定的奇素数 n 有关). 如同高斯整数或上面出现过的形如 $a+b\sqrt{-3}$ 这样的数，分圆整数也构成一个数系，它在某种程度上类似于普通的整数(它们组成一个环；相关的定义参见第 3 章).

1847 年 3 月 1 日，极度兴奋的拉梅向巴黎科学院的成员作报告，他说他终于证明了费马最后定理. 他的关键思想是利用了(现在所称的)分圆整数. 这使他通过无限递降法得到了完全的证明，它跟欧拉对 $n=3$ 时的论证十分相像.(他的论证中最关键的一步是证明了：若 x^n+y^n 的因子 $x+y,x+yr,\cdots,x+yr^{n-1}$ 无公因子，则 x^n+y^n 和 z^n 相等这一事实蕴含了每个因子 $x+y,x+yr,\cdots,x+yr^{n-1}$ 都必定是某个 n 次幂.)

讲完他所寻觅到的证明，拉梅感谢道，按上述方式使用复数是他

的同事J·刘维尔(Joseph Liouville)在几个月前向他建议的.当拉梅入座时,正好也在场的刘维尔问道,若仅仅证明了这些因子中没有两个有公因子,拉梅真的能推出 x^n+y^n 的每个因子都是 n 次幂吗?他接着指出,对于普通的整数这一推理是对的,但它依赖于唯一因子分解定理,而他知道对于分圆整数并不存在这样的定理. [191]

刘维尔事先是否知道欧拉先前犯的错误,我们不得而知.无论如何,他的发言正中拉梅论证的要害.悲哀而窘迫的拉梅经过几周勇敢的拼搏,试图拯救他的证明,但最后认识到他犯了个无法无天的错误.他写信给他在柏林的朋友狄利克雷:"如果你在巴黎或我在柏林,这一切可能不会发生."事实上,造成拉梅极度窘迫的局面也许可以避免,假如他知道E·库默尔(Ernst Eduard Kummer)约在3年前发表的一篇文章的话.(不知何故,库默尔选择了一份极不引人注目的刊物——《布雷斯劳大学庆祝哥尼斯堡大学周年纪念文集》,让他的这篇文章在那里沉睡,当然这对拉梅并不失公平.)

库默尔的工作和理想数

库默尔在1844年的文章中已证明,对分圆整数而言,唯一因子分解定理通常不成立,这一论断完全摧毁了拉梅对最后定理的证明.但是到1847年,当拉梅和数学界的其他人知道这一结论时,库默尔已研究出一套令人难忘的新理论,它表明有一种方法可以改变唯一因子分解的概念,以便建立适合于分圆整数的一种合理的"数论".他的理论的基础是在分圆整数的算术中引进他称之为理想素因子的概念——这一做法有点类似于在普通整数的算术中导入虚数 i.利用库默尔的理想数,涉及整数的唯一因子分解方面的许多结论可以对分圆域加以证明(对在证明费马最后定理的各种情形时出现的诸如 $a+b\sqrt{-3}$ 这样的数亦然).

库默尔的工作是自费马最后定理问世至(本章开头提及的)1983年那项成果出现前最重大的成就.他在1847年得到的成果已可对下 [192]

述情形证明最后定理:所有小于 37 的素指数(因而对所有小于 37 的指数亦然),除 37,59 和 67 以外的所有小于 100 的素指数. 这离数学家费尽心机得到 $n=5$ 和 $n=7$ 时的证明才刚过几年.

不仅如此,库默尔的关键性新概念理想数,原来还是一种特别有用和涉及范围极广的概念,它将引出一个更一般的概念——理想,以及一整个数学分支——理想论. 后者的基本原理现已成为大学数学系学生的例行课程. 虽然库默尔的理想数对费马最后定理的应用颇惹人注目,但事实证明,理想这一概念本身是库默尔对其他数学领域作出的最重要的贡献.

确实,正如库默尔这项成果的深远意义不在于对最后定理的应用一样,他的出发点本也不是为了证明费马的断言. 库默尔像高斯那样,一直试图将高斯证明了的二次互反律推广到高次互反律. 正是这种兴趣导致他于 1859 年证明了更强的一般性结论. 这件事告诫人们应注意费马最后定理跟高次互反律的密切联系.

正 则 素 数

库默尔的工作对最后定理的特殊价值在于,它提出了一种算术性质的条件,一种使最后定理成立的那些奇素数次幂必须满足的条件. 即若一个奇素数 p 满足库默尔的条件,则方程

$$x^p + y^p = z^p$$

无解. 今天,我们称满足库默尔条件的素数为**正则素数**. 小于 100 的素数中仅有 37,59 和 67 不是正则素数,这是库默尔本人于 1847 年证明的.

[193] 什么是正则素数呢? 此概念跟第 3 章描述的类数有密切联系. 正则素数是指不能整除相关分圆数域的类数的素数. 解释这个定义需要很多笔墨,幸好有另一个只涉及比较简单的概念的(等价)定义. 回忆第 3 章的内容,我们知道 e 是一个标准的数学常数,它用无穷小数表示时的头几位数字是 2.71828…,对于任何一数 t,e^t 的值由下

列无穷和给出：

$$e^t = 1 + \frac{t}{1!} + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots.$$

贝努里数 B_k 定义为下述无穷和的系数：

$$\frac{t}{e^t - 1} = 1 + B_1 \frac{t}{1!} + B_2 \frac{t^2}{2!} + B_3 \frac{t^3}{3!} + \cdots.$$

这些贝努里数的值的变化极无规则.对所有奇数 k（除了 $k = 1$，此时 $B_1 = -\frac{1}{2}$），B_k 为零.对偶数 k，前几个 B_k 的值为

$$B_2 = \frac{1}{6}, B_4 = -\frac{1}{30}, B_6 = \frac{1}{42}, B_8 = -\frac{1}{30},$$

$$B_{10} = \frac{5}{66}, B_{12} = -\frac{691}{2730}, B_{14} = \frac{7}{6}, B_{16} = -\frac{3617}{510}.$$

依据这些贝努里数，我们可定义正则素数：若素数 p 不能整除任一个 $B_2, B_4, \cdots, B_{p-3}$ 的分子，则它是正则的.（由此可知，若 p 至少能整除贝努里数 B_2、$B_4, \cdots, B_{p-3}$ 中的一个的分子，则 p 就是非正则的.）

贝努里数给出的正则性定义，提供了一种使用计算机检验素数正则性的方法；虽然直接按照上述定义进行检验的效率不高，但在实际应用时，可根据具体的贝努里数的特性导出更简易的计算方法.（检验正则性时，计算者面临的难题是有些贝努里数的分子非常大，[194] 例如

$$B_{34} = \frac{2577687858367}{6},$$

它仍可用手算来处理，而 B_{220} 的分子竟有 250 位数字.）

如上所述，最早从事判定正则与非正则素数的计算工作的是库默尔本人.他检验到 164，小于此数的非正则素数有 37，59，67，101，103、131，149 和 157.（对于每一种情形，库默尔都必须证明所讨论的素数能整除某个贝努里数的分子.例如，37 能整除 B_{32} 的分子，59 能整除 B_{44} 的分子，而 157 能整除 B_{62} 和 B_{110} 的分子.）

在 20 世纪 30 年代，斯塔福德（Stafford）和范迪弗（Vandiver）使用

台式计算机(并利用某些判定正则与非正则性的新方法)检验了直到617的所有素数.在1954年,随着电子计算机的出现,莱默(Lehmer)和范迪弗算到了4001,后来又有人算到30000.1976年,美国依利诺斯大学的S·瓦格斯塔夫(Wagstaff)借助于IBM 360-65和IBM 370计算机,对所有小于125000的素数的正则性作出了判断.

基于数值计算可知,大约60%的素数是正则的.为精确起见,对于较大的N,可得到如下比值:

$$\frac{\text{小于 } N \text{ 的非正则素数的个数}}{\text{小于 } N \text{ 的素数的个数}} \approx 0.39.$$

西格尔(Siegel)在1964年得到的虽不严格但可能成立的论证表明,上述的比值"应该是"$1-1/\sqrt{e}$;若取两位小数,其值恰为(我们所希望的)0.39.

尽管正则素数在数量上占有明显的优势,但我们并不能肯定它是否有无穷多个.令人惊奇的是,1915年詹森(Jensen)通过数值计算证明存在无穷多个非正则素数.这似乎是在说一个较大的集可能是[195] 有限的,而那个显然小一些的集倒已被证明是无限的!

现　　状

库默尔1847年的结论表明,若p是正则奇素数,则最后定理对幂指数p成立.但是p为非正则素数时如何呢?库默尔的结论对此问题无能为力.(显然并不能推出此时最后定理不成立,我们只能说库默尔的特殊论证方法此时不起作用,而定理本身仍可能由其他途径证明是正确的.)几年后,库默尔对素数附加了比正则性更一般(尽管不够精密)的条件,它也能保证最后定理成立.非正则素数37,59,67就满足这种条件,因此他恰好能宣告对于100以内的指数,最后定理都成立.此后人们发现了(附加于素数的)更广泛的条件,致使瓦格斯塔夫于1976年用计算机证明了,对所有小于125000的幂指数,最后定理都是正确的.

目前可搜集到的有关最后定理的其他信息，涉及上面提到过的将每种情形分解成两种从属的情形．对一给定的奇素数 p，第一种从属情形说，方程

$$x^p + y^p = z^p$$

不存在这样的本原解 x,y,z，使得 p 不能整除 x,y 和 z 中的任何一个；第二种从属情形说，不存在这样的本原解 x,y,z，使得 p 能整除 x,y 和 z 中的一个．经过若干年的努力，对第一种从属情形取得了一个重要的进步．

1832 年（比拉梅和库默尔的工作早许多年），法国数学家 S·热尔曼(Sophie Germain)证明：若 p 是奇素数并使得 $2p+1$ 也是素数，则最后定理的第一种从属情形对 p 成立（当时是这样说的：要是指数为 p 的费马方程有解，p 必须能整除组成该解的三个数中的一个）．尽管存在许多素数 p 使得 $2p+1$ 也是素数（例如 $p=3,5,11$），因而可以应用热尔曼的结论，但我们不知道是否存在无穷多个这样的素数．

接着，勒让德(Legendre)拓展了热尔曼的思想，对具有如下性质

所有素数成立.

1985年,美国人阿德勒曼(Adleman),法国人福里(Fouvry),英国人希斯-布朗(Heath-Brown)利用热尔曼准则的一种推广形式,第一次证明最后定理的第一种从属情形对无穷多个素数成立(尽管取得了这一进步,但最后定理仍可能仅对有限多个幂次的情形成立).

本章开头提到的法尔廷斯于1983年获得的引人注目的成果是什么呢?他证明,对每一个大于2的指数 n,费马方程

$$x^n + y^n = z^n$$

至多有有限多个本原解.这一证明使他赢得了1986年的菲尔兹(Fields)奖章.

它是否能导致最后定理的完全证明仍是个谜;但它把存在无穷多个解的可能性降到了最多只可能有有限多个解,这确实是迈出了一大步.(请注意,法尔廷斯的结论只是说解的个数有限,但没有指出解的个数的上限是多少.)

事实上,上述结论是法尔廷斯证明的更一般的结论(称为莫代尔猜想)的一种特殊情形.1922年,L·莫代尔(Lewis Mordell)猜测:任一系数为有理数的两变元不可约多项式,若亏格大于或等于2,那么它最多有有限多个(有理数)解.(如果你曾见到过"亏格"这个数学词汇,你就知道莫代尔猜想属于拓扑学的范围,我们将在第10章讨论[197]这门学问.)因为多项式

$$x^n + y^n = 1 \tag{7}$$

($n \geq 3$)满足莫代尔猜想中的假定,所以立即可知这个方程最多有有限个有理数解.由于方程

$$x^n + y^n = z^n \tag{8}$$

的任一整数解可导出方程(7)的有理数解(用 z^n 除方程(8)的两边),故(8)的不同的本原解将给出方程(7)的不同的有理数解,由此即可推出方程(8)仅有有限个本原整数解.

未　来

我们现在面临着什么样的形势呢？对直至125000的所有指数，费马最后定理都是对的；但除了附加了极强限制的第一种从属情形外，我们仍不知道是否有无穷多个指数使最后定理成立．超出我们已掌握的这个极限，也许有一个或许多素指数 p（比125000大）使该定理不成立；对这种 p，可能仅有有限个本原解存在．若 p 小于6000000000，那么（因这样的 p 使第一种从属情形成立），在任一解的三个数中至少有一个能被 p 整除，所以方程中会出现超过 125000^{125000} 的数．另一方面，若 p 比6000000000大，那么我们将遇到更大的天文数字．就任何实用的目的而言，最后定理是“真的”．当然，对数学家而言，事情绝不能就此结束．只有得到严格的证明或否证，最后定理的问题才算解决．目前，我们还不清楚已有的知识对达到这一目标有何用处．也许需要全新的方法，这样，最后定理又可能引导出对其他数学领域有重要意义的成果．

非常可能发生的情况（尽管不能绝对肯定）是，若想找到完全的解答，我们需要大大超出“初等”范围的知识．这意味着，众多的业余数学家（像费马？）——他们常常宣布自己证明了最后定理——几乎无例外地都犯了错误．事实上，仔细检查这类证明的话，通常总能发现他们的论证连 $n=3$ 的情形都证明不了，而那是欧拉在1753年就证明了的．人们发现，每年都有大量这类“证明”出现，错误“证明”中的绝大多数都要求得到使最后定理成为真正的定理的荣誉． [198]

大量此类“证明”都寄往西德哥廷根大学的数学研究所．第一个证明费马最后定理的人，将赢得沃尔夫斯克尔（Wolfskell）奖，并将获得法国科学院于1816年设的金质奖章和3000法郎奖金．前一项奖是哥廷根皇家科学院根据P·沃尔夫斯克尔的遗嘱设立的，1908年此奖初设时的奖金额是10万马克．此后，随着德国通货不断的常常还是令人吃惊的变化，目前的奖金额为1万多一点西德马克．（如果到

2007年9月13日仍无人能获得它,此项奖金将被取消.)

尽管对此奖的申请论文有许多严格的规定,哥廷根数学研究所仍然平均每周收到一篇应征论文要求评审.这比设奖头一年的情况好多了,那年共收到621份申请!

未受过训练的业余爱好者(且不说有经验的专家)获得成功的机会微乎其微.正因为如此,数学家并不鼓励别人来一试身手.由于做数学研究首先是兴趣所致,谁又会去干扰他人的快事呢?当你试着去做了,而且(极大的可能是)失败了,那么,你至少能因历史上众多最著名的数学家也未能证明这个撩人的问题而得到安慰;如果你去做了,而且成功了呢……

阅 读 文 献

至少有两本书出色而全面地介绍了有关费马最后定理的各种信息.Harold M.Edwards的Fermat's Last Theorem (Springer-Verlag,1977)中有3章的内容是非常"初等的",然后才转入较难懂的概念,后者体现了本世纪研究这一问题的基本特征. [199]

Paulo Ribenboim的13 Lectures on Fermat's Last Theorem(Springer-Verlag,1980)虽然没有给读者提供较易读的入门章节,但也出色地讲述了大量信息.

I.N.Stewart和D.O.Tall的Algebraic Number Theory(Chapman and Hall,1979)涉及相当多的有关费马最后定理的内容和相关的课题,此书的对象是大学数学系的学生.

类似程度的另一本很出色的读本是David Burton的Elementary Number Theory(Allyn and Bacon,1980),它不仅讨论了费马最后定理,还讲述了数论中所有其他的课题. [200]

(袁向东译)

译后记:

费马最后定理经过三个多世纪的漫长历程,终于在1994年9月得到了完

全的证明．最后的攻坚路线跟费马本人、欧拉和库默尔等人的完全不同，它是现代数学许多分支(诸如椭圆曲线理论，模形式理论，伽罗华表示理论，等等)综合发挥作用的结果．由于整个证明过程涉及众多高深的数学理论，许许多多数学家为此作出了贡献，我们无法在这里一一细述，只能极粗略地勾划出证明路线的轮廓．

在本世纪50～60年代，数论研究中逐渐形成了一个重要猜想，它最早由谷山丰(Y，Taniyama)提出，后经志村五郎(Goro Shimura)和A·韦尔(Weil)精炼成如下形式：有理数域上的每条椭圆曲线都是模曲线．(现在一般称之为谷山－志村猜想．)

从60年代后期开始，有人将费马方程$x^n+y^n=z^n$和形如

$$y^2=x(x+A)(x+B) \tag{1}$$

的椭圆曲线相联系，最初的着眼点是利用跟费马最后定理有关的结论来证明与椭圆曲线有关的结论．1985年，G·弗赖(Frey)在两者的联系方面迈出了关键的一步，他在一次演说中提出：假定费马最后定理不成立，即存在一组非零整数a,b,c，使得$a^n+b^n=c^n(n>2)$，那么用这组解构造出的形如(1)的椭圆曲线(即在(1)中取$A=a^n$，$B=-b^n$，现称这类椭圆曲线为弗赖曲线)，不可能是模曲线．此结论显然和谷山－志村猜想矛盾．如果弗赖的结论和志村猜想都得到证明是正确的，根据反证法的逻辑可知，“假定费马最后定理不成立”必是错的，因而导出费马最后定理成立．可惜弗赖本人未能证明他的结论；1986年，K·里贝(Ribet)按照J·－P·塞尔(Serre)的思想证明了弗赖的论断．于是，证明费马最后定理的工作归结为去证明谷山－志村猜想．

当时的数学家们普遍认为，要证明谷山－志村猜想还是很遥远的事．但是英国数学家A·怀尔斯(Wiles)对这种看法不以为然，他立即集中全副精力去证明这个猜想．经过7年的奋斗，怀尔斯于1993年6月在英国剑桥大学牛顿数学科学研究所举行的数学讨论会上，报告了他对如下结论的证明：对有理数域上一大类椭圆曲线(用专业术语应称为半稳定的椭圆曲线)，谷山－志村猜想成立．由于弗赖曲线恰好属于半稳定的椭圆曲线的范围，因此费马最后定理自然成为怀尔斯的结果的推论．据称怀尔斯的证明长达200页．按照数学界的习惯，他的证明在得到最后确认前，必须经过其他有关数学家的详细审查，尽管当时许多人相信怀尔斯的证明是经得起推敲的．好事多磨，事情并未就此了结．有关怀尔斯的证明中存在漏洞的传闻不胫而走．1993年12月4日，怀尔斯向同行们发出一份电子邮件，承认他的证明中确有漏洞．数学家对待证明的态度是十分

严肃的，不容半点含混.1994年10月25日，美国俄亥俄州立大学教授K·鲁宾(Rubin)以电子邮件向数学界的朋友发出了谨慎而乐观的消息：

“今天早晨，有两篇论文已经公开，它们是：‘模椭圆曲线和费马最后定理’，作者是A·怀尔斯；‘某些赫克(Hecke)代数的环论性质’，作者是R·泰勒(Taylor)和A·怀尔斯.第一篇是一篇长文，……它宣布了费马最后定理的一个证明，而这个证明中关键的一步依赖于第二篇短文.……怀尔斯在他的剑桥演说中所描述的证明被发现有严重的漏洞，它涉及欧拉系的构造.在怀尔斯努力补救这个构造未获成功之后，他回到他原先曾试过的另一途径，以前由于他偏爱欧拉系的想法而放弃了这个途径.在作了某些赫克代数是局部完全交的假设之后，他可以完成他的证明.这一想法以及怀尔斯在剑桥演讲中描述的其余想法，写成了第一篇论文.怀尔斯和泰勒合作，在第二篇论文中建立了所需的赫克代数的性质.

证明的整个纲要和怀尔斯在剑桥描述的那个相似.新的证明由于排除了欧拉系，比原来的那个更简单和简短了.……虽然在稍长一点时间里保持小心谨慎是明智的，但是肯定有理由表示乐观.”

1995年7月号的“美国数学会通告”上刊出了G·法尔廷斯(他就是本章开头提到的那位著名人物)的文章，题为“R·泰勒和A·怀尔斯对费马最后定理的证明”.他开宗明义，以肯定的语调宣称：“在本文题目中所提到的猜想于1994年9月终于被完全证明了.”至此，人们相信那个搅扰了数学家300多年的著名猜想真正成了一条定理！

第9章 复数领域的难题

一个复杂的领域

对许多读者,这将是本书最困难的章节.这里的数学内容本身并不比前面的难,但其抽象程度却增加了.数——自然数和复数——构成了这一领域的核心.然而*复分析*(与*复函数论*一词几乎同义)和与它密切联系的*解析数论*(复分析的结果和方法在自然数研究中的应用)的基本任务是要发现和探索看来相当简单的观念(见第3章中有关复数的描述)背后的深层结构和相互关系.进行这一领域的研究要用到数学界外的许多人所不熟悉的非常抽象的数学方法.而且不幸的是我们不能借助图形来理解这一课题,因为它不形象化(在第10章中同样困难和陌生的拓扑学概念至少可以通过简单的例子用图表表达).但这确是一个重要的领域,近些年来取得了一些显著的进步,所以是不应被忽视的.还有,如果你坚持读完这一章就会发现,透过抽象的内容,你对诸如分数和素数这样熟悉的概念又有了更深的领悟(假设你已读过第3章中介绍复数的段落).

201]

虽然构成这一章核心的三个问题都是高度抽象的,但并不是说复分析在数学之外没有用处.远非如此,从1825年柯西(Augustin Cauchy)在这一课题上的最初工作开始,它就一直都同外部世界有联系.柯西之后,黎曼(Riemann)在此课题上的工作显示了复分析是怎样极大地帮助了物理问题的解决,在所谓"积分变换"(如无处不在的傅立叶(Fourier)变换)上的进一步工作使这种联系更加明显.

复数的二维性质（参看第3章对复平面的讨论）使它能被用来研究二维的问题，其方式如同用实数来处理一维问题．因三维空间中的实际生活问题有对称的性质（比如通过圆管的液体流），所以常被简化为二维的数学问题．这样复分析就同物理学家和工程师有关．

俄国数学家儒可夫斯基(N.Y.Joukowski,1847～1921)曾用复分析来确定机翼的形状(即机翼的截面形状),并研究它周围空气流的类型,这给飞机的设计带来了革命性的变化;从此以后,复变函数理论在描述各种流体流动的形式及车船设计上占据了重要的中心位置.1920年,美国贝尔实验室的科学家们将复变函数理论系统地应用在滤波器和高倍放大器的设计上,它使远距离的电话通信成为可能(如果你认识一位电子工程师,可以问问他判定反馈放大器稳定性的内魁斯特(Nyquist)准则的重要性,这是复分析的一个直接应用).简而言之,复分析现在是一个不可或缺的工具,在某种程度上,今日的科学和技术很少有不依赖于复数的.

那么复分析是什么样的学科呢？我们先尝试给出一个回答:微积分方法(微分、积分、无穷和等等)从我们熟悉的实数到复数领域的扩展.但正如所有类似的尝试一样,这个回答既有启发性,又有误导性,因为当微积分的全部工具被扩展到复数领域时,具有了完全不同的形式.实数中似乎截然不同的概念,在引入复数以后可能会联系得非常紧密. [202]

这方面的一个例子已在第3章提过,就是欧拉的恒等式:

$$e^{\pi i} = -1.$$

与它紧密联系的另一个例子是方程式

$$e^{xi} = \cos x + (\sin x)i$$

(x 为任意实数),其中涉及到 e,i 和我们熟悉的正弦和余弦三角函数.事实上,把正弦和余弦函数用于复数并没有限制.当然你无法进行计算,比如 $\sin(3+4i)$就不能像实数那样用直角三角形来计算.你一旦要和复数打交道,就要做好被理论引导到任何地方的准备.对三

角函数来说，将被引向无穷级数的领域. 比如无限和

$$e^{x}=1+\frac{x}{1!}+\frac{x^{2}}{2!}+\frac{x^{3}}{3!}+\cdots$$

不管 x 是实数还是复数都给出一个有效的答数，无穷级数方程

$$\sin x=x-\frac{x^{3}}{3!}+\frac{x^{5}}{5!}-\frac{x^{7}}{7!}+\cdots,$$

$$\cos x=1-\frac{x^{2}}{2!}+\frac{x^{4}}{4!}-\frac{x^{6}}{6!}+\cdots$$

也是这样. 如果 x 为实数，每个无限和都会准确地给出与在通常的几何定义下完全一样的答案（“角”x 不是角度而是弧度，1 弧度等于 $\frac{180}{\pi}$ 度，即约 57.3 度）. 但 x 如果为复数，并没有什么理由可以阻止你使用这些方程式.

对以上三个表达式进行简单的代数运算，可以进一步得到：

$$\sin x=\frac{1}{2i}(e^{xi}-e^{-xi}),\cos x=\frac{1}{2}(e^{xi}+e^{-xi}),$$

[203] x 为实数或复数均可.

复分析的首要任务之一是检查所有以上的涉及无穷级数的运算是否能被允许. 第 2 章已清楚地阐明，无穷概念必须小心处理，特别是其中涉及复数时就更是如此.

复变函数的积分[①]与实积分明显不同. 因为复数是二维的，所以复数当然不能像实数那样从一个数 a 积分到另一个数 b，比如像：

$$\int_{0}^{1}x^{2}\mathrm{d}x=\frac{1}{3}.$$

相反你必须沿着（复平面内的）某一条路径积分. 例如你也许想环绕一个圆周对某一函数实施积分. 比如说（复变）函数 $\frac{1}{(x-a)}$（a 为一复常数）沿着圆形道路 C 积分的结果怎样呢？（这个积分写成

---

① 如果你不熟悉实数的微积分概念，你可以跳过这一部分的其余内容和此章中其他偶尔出现的有关积分的参考材料. ——原注.

$$\int_C \frac{1}{x-a}\mathrm{d}x.$$

这种积分有时指线积分.)积分结果相当出人意料,对那些还记得实函数的积分是多么困难的读者来说尤其如此.如果数 a 对应于复平面上在圆 C 内的一点,结果是 $2\pi i$;如果数 a 在圆 C 外,结果是 0.奇怪的是结果与圆的大小和位置一点没有关系,常量 a 影响结果的唯一途径是它落在圆 C 内部或外面(虽然选择这个特例是为了提供这样一个惊人的结果,但它同时也典型地说明了复变函数有其自身的生命力,远远超出了我们从实数而来的想象.)

把复数理论应用于自然数的研究会得到更多的令人惊讶的结果,这一点随着此章的展开会表现得很清楚.(值得注意的是以上提到的积分在这方面的研究中起着重要作用,虽然涉及的不是这里所能解释的那种积分.)

[204]

一些数字游戏

一个分数 h/k,当其值在 0 和 1 之间,且 h 与 k 没有公共因子时叫做真分数.例如:1/2,3/4 和 7/8 都是真分数;而 2/4,3/9,3/2 就不是.对任何数 n,由分母不超过 n 的真分数,连同分数 1/1 按递增顺序组成的数列 F_n 叫做法里(Farey)数列.所以 F_5 是数列:

$$\frac{1}{5},\frac{1}{4},\frac{1}{3},\frac{2}{5},\frac{1}{2},\frac{3}{5},\frac{2}{3},\frac{3}{4},\frac{4}{5},\frac{1}{1}.$$

F_7 就是数列:

$$\frac{1}{7},\frac{1}{6},\frac{1}{5},\frac{1}{4},\frac{2}{7},\frac{1}{3},\frac{2}{5},\frac{3}{7},\frac{1}{2},\frac{4}{7},\frac{3}{5},\frac{2}{3},\frac{5}{7},\frac{3}{4},\frac{4}{5},\frac{5}{6},\frac{6}{7},\frac{1}{1}.$$

不知道是谁首先想到要研究这种数列的,但第一个天才地证明了这个数列的规律的人似乎是哈罗斯(Haros),时间是 1802 年.1816 年,法里在一篇文章中陈述了哈罗斯的结果但没有给出证明,后来柯西看到这篇文章,他发现了这个结果的一个证明,并把这一概念的产生归功于法里,所以这个数列就叫做"法里数列".

那个先被哈罗斯证明后由法里陈述的结果是：从一个法里数列中取出任意相邻的三项：a/d，b/e，c/f，则 $b/e=(a+c)/(d+f)$。例如，F_7 的第10，第11和第12项是4/7，3/5和2/3，而

$$\frac{4+2}{7+3}=\frac{6}{10}=\frac{3}{5}.$$

哈罗斯还证明了另一个结果：如果 a/c，b/d 是相邻的法里数列的项，则 $bc-ad=1$。比如 F_7 的第6项和第7项是1/3和2/5，则

205」

$$2\times3-1\times5=6-5=1.$$

上面的两个陈述可以互相推出，就是说为证明两者正确，你只需证其中的一个。两者都可以作为有意义的代数运算练习。（如这不对你的口味，你至少可以对其他某几个法里数列来检验上述结论。）

对任意数 n，让 $A(n)$ 代表法里数列 F_n 的项数，这样 $A(5)=10$，$A(7)=18$。假设取实轴上从0到1的一段并将其分成相等的 $A(n)$ 段（图47）。分点为 $1/A(n)$，$2/A(n)$，$3/A(n)$，直到 $(A(n)-1)/A(n)$。因为法里数列 F_n 的项在0到1之间分布不匀，数列中的许多项的值不会碰巧都与分点相合。令 d_1 为法里数列的第1项与 $1/A(n)$ 的差，d_2 为第2项与 $2/A(n)$ 的差，如此到 $d_{A(n)-1}$。（两个数中哪个大并不要紧，关键的是它们的差。）令 $D(n)$ 为 $d_1, d_2, \cdots, d_{A(n)-1}$ 的和。

例如，F_4 由

$$\frac{1}{4},\frac{1}{3},\frac{1}{2},\frac{2}{3},\frac{3}{4},\frac{1}{1}$$

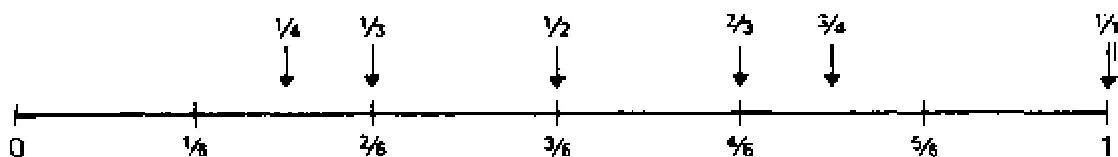


图47 法里数列 F_4。这个数列的成员由箭头标出，并示意了它们与将0到1区间分成相等6份的5个等分点的相对位置。数 $d_1, d_2, \cdots, d_5$ 表示法里分数与相应的等分点的差，$D(4)$ 是这些差的和。

组成. 故 $A(4)=6$. 把 0 和 1 间的距离平分成 6 小段的点为 1/6,2/6,3/6,4/6,5/6. 故 d_1 是 1/4 与 1/6 的差, $1/4-1/6=1/12$; d_2 是 1/3 与 2/6 的差,即 0; d_3 是 1/2 与 3/6 的差,也是 0;同样 $d_4=0$; $d_5=5/6-3/4=1/12$. 故 [206]

$$D(4)=\frac{1}{12}+0+0+0+\frac{1}{12}=\frac{1}{6}.$$

(看到这里你应自己试试算出 $D(5)$的值.)

在 1942 年的一篇论文中,弗兰内尔(J. Franel)和兰道(E. Landau)探讨了当 n 跑遍所有自然数时函数 $D(n)$的表现.(他们用的是代数工具,而不是用算术方法对很多法里分数进行计算.)特别地,他们考虑了命题:如果 r 是任一大于 1/2 的实数,则存在一个常数 C,使 $D(n)$总(即对任意的 n)小于 Cn^r. 他们证明的这个看似简单的命题是与当今职业数学家普遍认为的这一领域内最重要的未解决问题相等价的(即是后者的另一种表述方式),这个未解问题就是:黎曼假设.

数学中最重要的未解决问题

对外行来讲,数学中最著名的未解决问题无疑是费马最后定理,有关内容在第 8 章已作了介绍. 但著名的未必是重要的. 如果问一位职业数学家,整个领域中唯一最重要的未解决问题是什么,你肯定会得到这样的回答:"黎曼假设". 英国伟大的数学家哈代(G. H. Hardy,见第 4 章)显然是这样想的. 一次他要从斯堪的纳维亚渡海到英格兰,临行时北海的天气出奇地糟糕,他在给一位同事的明信片上写道(无疑他脑子里想到了费马最后定理的产生):"已经证明了黎曼假设,你的哈代."哈代的意思是说,不证明这么重要的结果,上帝是不会让他死去的,所以一定会保佑他平安回家. 后来哈代安全地返回了(他是个地道的无神论者!),但"哈代大定理"却没有叫响. 直至今日黎曼假设仍未证明. [207]

黎曼假设的故事开始于1740年，当时欧拉引入了所谓ζ函数：若s为大于1的实数，定义：

$$\zeta(s)=\frac{1}{1^s}+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{4^s}+\cdots$$

（ζ是一个希腊字母，读作"采他"——zeta．"ζ函数"这个名称除了因为是用ζ这个字母来标记外，别无其他意义．）若s小于或等于1时，无限和结果为无穷，这种情况下，ζ函数没有意义．而对任意大于1的s，无限和有确定的有限的值．欧拉证明了对任意这样的s，$\zeta(s)$的值等于无穷乘积

$$\frac{1}{1-(1/2)^s}\times\frac{1}{1-(1/3)^s}\times\frac{1}{1-(1/5)^s}\times\frac{1}{1-(1/7)^s}\times\frac{1}{1-(1/11)^s}\times\cdots,$$

这里每个因子形如：

$$\frac{1}{1-(1/p)^s},$$

p是一素数．这个结果有两点是令人吃惊的．首先，它显示ζ函数与基本的素数是紧密联系的；其次，两者间的关系在很大程度上涉及到无穷．显然这已超出了素数给我们的感觉．

在故事的下一段，我们暂时从ζ函数转到素数分布这个问题上来．当依次地检查自然数时，一开始素数出现得较频繁（比如从1到10的自然数中一半是素数），但往后则变得稀疏．在余下的数中，素数分布似乎又没有统一的模式，如在9 999 900到10 000 000中间有9个素数，但在接着的100个整数10 000 000到10 000 100中只有两个素数：10 000 019和10 000 079．实际上，存在着不包含任何素数的整数区间．（容易看出：若N为任意值，则$N!+2$到$N!+N$的区间不包括任何素数．）素数出现的这种明显的随机性的另一个例子是许多对"孪生素数"的存在．孪生素数是相差为2的素数，如3和5，11和13，10006427和10006429，它们的出现好像也是随意的．（据猜测有无穷多对孪生素数，但没有被证明．）

[208]

但是，在表面上的混乱背后，素数的出现仍有一定规律，它涉及

到函数 $\pi(n)$的行为. $\pi(n)$告诉你小于或等于 n 的素数的个数(见第1章). 勒让德在他的《数论》(Essai sur la Théorie des Nombres, 1798)一书中注意到 $\pi(n)$近似等于:

$$\frac{n}{\log_e n - 1.08366}$$

(本章中 $\log_e n$ 表示 n 的自然对数,即以 e 为底.)此处的 1.08366 其实并不奇特. 勒让德是在检查了 400 000 以内的素数表之后得到其结果的,选择该数不过是为了给出尽可能好的结果.

大约在勒让德写他那本书的同时,14 岁的高斯也开始了对函数 $\pi(n)$的研究. 他发现(迟至 1863 年才发表) $\pi(n)$可由 $n/\log_e n$ 近似,还可由

$$\mathrm{Li}(n) = \int_2^n \frac{1}{\log_e x}\mathrm{d}x$$

近似.(Li 函数是"对数积分"函数. 读者如果不明白它的意义也不要紧;只是为了求"全"我们才将它包括进来.)表 2 给出了 n 取直到 100 000 000 时各近似函数的值. 从此表可看出 Li(n)远比其他两种函数更近似于 $\pi(n)$,事实上在 1896 年瓦莱·普桑(Charles de la Vallée Poussin)说明了从某点开始,对所有 n 值,Li(n)确是最接近于 $\pi(n)$的.

说来有点离题,但这也是有趣的,从表 2 看来,Li(n)总是要比 $\pi(n)$大一些,如果继续计算列表的话,那么我们观察到的现象似乎总是这样的. 要是由此还不能得出结论:Li(n)总是在比 $\pi(n)$大的一边近似于它,岂不是太多疑了? 但假如真的作出这种判断那就错了! 1914 年,英国数学家利特伍德(J. E. Littlewood,哈代的同事)说明了 Li(n) - $\pi(n)$的差随 n 跑遍正整数时在正负之间变化无数次. 所以肯定存在 n 的某个值,使 Li(n)小于 $\pi(n)$. 事实上,1955 年斯古斯(S. Skewes)证明了这样的 n 必在数

[209]

$$e^{e^{e^{79}}} \quad (约\ 10^{10^{10^{34}}})$$

之前某处出现. 这是个难以想象的大数,这个数也就叫做斯古斯数.

n	$\pi(n)$	$\frac{n}{\log_e n - 1.08366}$	$\frac{n}{\log_e n}$	Li(n)
1000	168	172	145	178
10000	1229	1231	1086	1246
100000	9592	9588	8686	9630
1000000	78498	78534	72382	78628
10000000	664579	665138	620420	664918
100000000	5761455	5769341	5428681	5762209

表 2 素数的分布．这是表 1(第 1 章)的扩展，它指出了对各种不同的 n 值比 n 小的素数的个数 $\pi(n)$，同时给出了三种经典的对 $\pi(n)$ 的近似函数的值．

1966 年勒曼(Lehman)发现了可以用 1.65×10^{1165}代替斯古斯数(即在此数之前某处 $\mathrm{Li}(n)-\pi(n)$要变号)．它比斯古斯数小得多，但仍是人类难以驾驭的数．1986 年，黎勒(H．J．J．te Riele)又找到了更小的数 6.69×10^{370}，即比它小的某数 n，将使 $\mathrm{Li}(n)-\pi(n)$符号改变．但这仍是个超巨型的大数，它使人们怀疑，可能永远都找不到使 $\mathrm{Li}(n)$小于 $\pi(n)$的某个确切的数 n(计算机搜索了十亿(10^9)以内的数，仍没有找到这样的 n)．[210]

现在我们回到故事的主线上来．1896 年，法国人阿达玛(J．Hadamard)和瓦莱·普桑独立地证明了已为各种迹象所提示的事实：当 n 增大时，$n/\log n$ 的值与 $\pi(n)$越来越近，不管你要这个近似值多么接近于 $\pi(n)$，只要不是相等，这种精确度总会由选择足够大的 n 而得到(当 n 变大时，$\mathrm{Li}(n)$当然也“任意地逼近”)．这个著名的定理叫做素数定理，它说明了素数的出现还是有一定的数学模式的．(不过，请注意素数背后的这种模式涉及无穷和微积分的一些概念．)这两个数学家的工作都是依赖于德国人黎曼(Bernhard Riemann)写于

1859年的一篇著名的8页的论文,文章可译为:《比给定量小的素数的个数》.黎曼在他这篇唯一的数论文章中鼓动人们在某些方向上进行研究.直到今天,事实证明,这些研究方向仍是非常富有成果的.事实上,他的文章的发表标志着整个解析数论领域研究的开始,微积分的有力方法被应用到自然数问题的研究上.

黎曼在其素数分布的研究中引入的重点概念是 ζ 函数的推广,s 不仅限于大于1的实数,而且可以是任意复数 $s=a+bi$(但 $s\neq 1$).这不能简单地让 s 在原来欧拉的定义

$$\zeta(s)=\frac{1}{1^s}+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{4^s}+\cdots$$

中取复数,而要用一种叫做解析延拓的相当复杂的技术(这里不作描述).为方便那些能理解所涉及到的各种符号(无疑会使剩下的人感到困惑)的读者,此处写出黎曼得到的推广 ζ 函数公式,它是用线积分方式表出的:

$$\zeta(s)=\frac{\prod(-s)}{2\pi i}\int_C\frac{(-x)^s\mathrm{d}x}{e^x-1\quad x}.$$

积分路径 C 从 ∞ 沿正实轴向左,止于原点附近,然后沿逆时针方向的圆绕过原点,再沿正实轴回到 ∞.乘积 $\prod(s)$ 为 [211]

$$\prod(s)=\lim_{N\to\infty}\frac{N!}{(s+1)(s+2)\cdots(s+N)}(N+1)^s,$$

这里 s 可以是所有不为负整数的数.

这个扩展的函数叫做黎曼 ζ 函数.事实证明,它在数论中有基础性的重要地位.有关这个问题出版了大量书籍.(其中有爱德华兹(*H.M.Edwards*)的300页的《黎曼 ζ 函数》,从上述推广的 ζ 函数的定义可知,这种书是仅供专家读的.)

对任意等于 $-2,-4,-6,\cdots$ 的 s,$\zeta(s)$ 值为0.也就是说负偶数是 ζ 函数的零点.此外,还有无数个其他的复数 s 使 $\zeta(s)=0$.其实数部分的值介于0和1之间(即他们有形式 $s=a+bi$,a 介于0和1之间).黎曼假设是黎曼在他的文章中对有关 ζ 函数的这些复零点

所作的猜想.他说(几乎没有任何根据)ζ 函数的所有复零点实数部分恰等于$\frac{1}{2}$(即,若 $\zeta(s)=0$,则 s 具有$\frac{1}{2}+bi$ 的形式).

为什么这个假说这么重要?正如黎曼在文中指出的,ζ 函数的零点和$\pi(n)$函数之间有着非常紧密的关系.正是这种联系使阿达玛和瓦莱·普桑作出了令人注目的素数定理的证明(他们证明的依据是其真实性已经确知的联系,而不是至今尚未获证的关于零点的黎曼假设).这种联系同样隐藏在大量的有关素数的其他已知事实的背后,包括前面提到的在斯古斯数上的工作.如果黎曼假设被证明是正确的,即 ζ 函数的零点真是那么有秩序的话,则 ζ 函数与$\pi(n)$函数的联系将使我们得出比现在所知多得多的有关素数的信息.这就是对数学家来说这个问题如此重要的原因.那么关于这个可能成立的假设我们已经知道些什么呢?

在描述黎曼假设提出以后在这方面所做的许多工作之前,也许 [212] 值得引述一个明确地涉及 ζ 函数的$\pi(n)$的某一特定近似函数.(只需要 ζ 函数的欧拉原始形式.)它是这样的:

$$R(n)=1+\sum_{k=1}^{\infty}\frac{1}{k\zeta(k+1)}\frac{(\log n)^k}{k!}.$$

(许多读者可能对符号不熟悉,别担心,读下去.表达式中的求和是个无穷求和——从 $k=1$ 到 $k=\infty$.)

表 3 表明 $R(n)$是对 $\pi(n)$的一个多么好的近似.

n	$\pi(n)$	$R(n)$
100000000	5761455	5761552
200000000	11078937	11079090
300000000	16252325	16252355
400000000	21336326	21336185
500000000	26355867	26355517
600000000	31324703	31324622

续表

n	$\pi(n)$	$R(n)$
700000000	36252931	36252719
800000000	41146179	41146248
900000000	46009215	46009949
1000000000	50847534	50847455

表3 素数的分布.此表紧接着表2结尾的数据,对这些较大的 n 值,用黎曼 ζ 函数定义的函数 $R(n)$ 给出了素数分布函数 $\pi(n)$ 的一个极好的近似.

[213]

黎曼假设

是什么证据支持黎曼的猜想:如果对一复数 s,$\zeta(s)=0$,那么 s 一定形如 $\frac{1}{2}+bi$ 呢?前面提到,任何的复零点其实数部分总在0和1之间,而 $\frac{1}{2}$ 正在中央,但黎曼一定有比这更充分的理由.不管他的主张的基础是什么,现在活在世界上的人无人知晓.我们不久就会清楚,黎曼并没有去一一计算零点,即使他那样做了,有关 $\mathrm{Li}(n)-\pi(n)$ 的符号的例子说明在数学中大量的例证也是靠不住的.G·H·哈代证明了(并不容易地)有无限多个零点的实部等于 $\frac{1}{2}$,但这并不能排除有无限多个更多的零点实部不为 $\frac{1}{2}$.除此之外,这一问题确实没有太大的进展.不过,多年的大量计算工作积累了庞大的证据.(虽然这些证据不能证明假设成立,但如果假设错误的话——虽不太像——总是可以通过大量的计算找到一个数作为反例的.要否定黎曼假设只须发现一个零点的实部不等于 $\frac{1}{2}$.这就是要进行计算的理由.另一个原因是这种计算为大量新算法提供了绝好的试验手段,而这

些新算法将来有可能有别的用处.）

有些算法能计算当复数$\frac{1}{2}+bi$是ζ函数的零点时的b值；还有些算法能计算虚部在任一区间上时零点的个数，把这两种算法结合起来就能够检查在给定的有限区间上黎曼假设是否正确.第一个设计了黎曼假设算法的人是格兰姆(J.－P.Gram).1903年，他用标准的欧拉－马克劳林求和法计算了零点.格兰姆得到了使$\zeta(\frac{1}{2}+bi)=0$的前15个b值.其中前10个他计算到了6位十进小数，第一个是$b=14.134725$，第10个是$b=49.773832$.剩下的5个他只计算到了小数点后一位，第11个是$b=52.8$，第15个是$b=65.0$.因知道复零点的实部在0和1之间，所以他接着指出，虚部在0和50之间的零点正好有10个.因为他早先列出的虚部在0与50之间形如$\frac{1}{2}+bi$的零点恰有10个，那么他的表恰包含了虚部在这一区间的所有零点.换句话说，他的计算证明了黎曼假设在0到50的区间上成立.(这里的“区间”指零点虚部的大小.)

〔214〕

用类似的方法(略有改进)，白克伦德(R.Backlund)在1918年证明了对0到200之间的零点，黎曼假设成立.1925年J·J·霍钦森(用进一步改进的方法)把上限提高到300.1936年，堤池马什(Titchmarsh)和孔瑞(Comrie)用卡尔·西格尔(Carl Siegel)设计的改进方法计算了1041个零点，都具有黎曼假设的形式.

二战以后，电子计算机被引入这一问题的研究.在50年代D·H·莱默(Lehmer)开始进一步的探索，他用西格尔的方法计算形如$\frac{1}{2}+bi$的零点，并结合了阿兰·图林(Alan Turing)提出的在一给定区间判定零点个数的新方法.1966年R·S·勒曼计算出了250 000个零点.以后几年里，罗塞尔(J.B.Rosser)和他的同事把零点个数提高到了3 500 000.1983年，在范德隆(J.van de Lune)和黎勒的工作之后，从0到119 590 809.282的区间都被检查过了，其中有300 000 001个零点，它们都是黎曼所预料的类型.1985年还是他们两个人计算了

前15亿个零点,还没有发现黎曼假设的反例.

这便是支持黎曼假设的全部数值上的例证.如果它不对,那么给出相悖结果的数必定是超出了职业纯数学家之外的人通常所考虑的范围.这一点上黎曼假设与费马最后定理(见第8章)真是一对难兄难弟.尽管有以上所有的工作和证据,还是无人能知道黎曼假设究竟对与不对.但似乎是再次警告那种试图在数值例证的基础上便归纳出结论的诱惑,在过去的几年当中,一个与黎曼假设密切相关的猜想已被解决.在这一情形中,事实证明数值例证最终会完全使人误入歧途.默顿斯猜想的命运应被当做是对认为黎曼猜想一定正确的人的警告. [215]

默 顿 斯 猜 想

任取自然数n,则由算术基本定理,n或者为一素数,或者可以唯一表示为一些素数的积.比如,对前5个非素数:

$$4=2\times2,\quad 6=2\times3,\quad 8=2\times2\times2,$$
$$9=3\times3,\quad 10=2\times5.$$

4,8,9等数的分解,至少有一个素数因子出现一次以上.而6和10的分解,每个素数只出现一次.可被一个素数的平方整除(如4,8,9)的数叫做平方可除数(square-divisible),不能这样整除的数叫做无平方因子数(square-free).(这样,一个无平方因子数的素数分解中,没有一个素因子出现一次以上.)

如果 n 是一非素的无平方因子的自然数,则它要么是奇数个素数的积,要么是偶数个素数的积.比如,$6=2\times3$ 是偶数个素数的积,$42=2\times3\times7$ 是奇数个素数的积.1832年,莫比乌斯(A.F.Möbius)引入了下面的简单的函数 $\mu(n)$(用希腊字母 μ 表示,并称作莫比乌斯函数)以表示数 n 的素因子分解类型.特别地,当 $n=1$ 时,令 $\mu(n)=1$;对其他所有的 n,$\mu(n)$定义如下:

若 n 平方可除,则 $\mu(n)=0$;

若 n 无平方因子,同时是偶数个素数的积,则 $\mu(n)=1$;

若 n 是素数；或 n 无平方因子，同时为奇数个素因子的积，则 $\mu(n) = -1$.

举例来说，$\mu(4) = 0$，$\mu(5) = -1$，$\mu(6) = 1$，$\mu(42) = -1$，你可计算更多的值.

现对任何数 n，令 $M(n)$表示对所有小于等于 n 的k 值的$\mu(k)$ [216] 的和，比如：

$$M(1) = \mu(1) = 1,$$
$$M(2) = \mu(1) + \mu(2) = 1 + (-1) = 0,$$
$$M(3) = \mu(1) + \mu(2) + \mu(3) = 1 + (-1) + (-1) = -1,$$
$$M(4) = \mu(1) + \mu(2) + \mu(3) + \mu(4) = 1 + (-1) + (-1) + 0 = -1,$$
$$M(5) = 1 + (-1) + (-1) + 0 + (-1) = -2,$$

读者可自行验证：

$$M(6) = -1,\quad M(7) = -2,\quad M(8) = -2,$$
$$M(9) = -2,\quad M(10) = -1,\quad M(11) = -2,$$
$$M(12) = -2,\quad M(13) = -3,\quad M(14) = -2,$$
$$M(15) = -1,\quad M(16) = -1,\quad M(17) = -2,$$
$$M(18) = -2,\quad M(19) = -3,\quad M(20) = -3.$$

(问题：使 $M(n)$重新为 0 或重新为正的第一个 n 值是多少?)

这看起来好像是一种令人愉快的游戏，很难让人联想到它与数学中最重要的未解问题有什么关系.(虽然读过法里数列的部分，你也许不是很肯定.)一会儿就会知道，函数 $M(n)$的行为与黎曼 ζ 函数的零点位置有着密切的关系.

斯蒂阶(T.J.Stieltjes)看来是知道这种关系的.1885 年，在给他的同事赫尔米特(C.Hermite)的一封信中，他声称已证明了：不论 n 多大，$M(n)$的绝对值总小于$\sqrt{n}$；即

$$|M(n)| < \sqrt{n}. \tag{9}$$

(两条竖线是一种标准记号，表示消去表达式内的任何负号.如

$|-10|=10$, $|5|=5$, 等等.)如果斯蒂阶所言正确,那么黎曼假设的正确性立即成立.事实上,如对任意的 n,存在一常数 A,使不等式 [217]

$$|M(n)|<A\sqrt{n}$$

成立,则黎曼假设成立.不用说,根据我们前面一节的观点,斯蒂阶错了,可在当时却一点都不清楚.(例如,当1896年阿达玛的那篇现已被奉为经典,并获得一致好评的证明素数定理的论文中提到:他懂得斯蒂阶已用不等式(9)得到了相同的结果,并辩解说因为斯蒂阶没有公开他的证明,所以自己才发表了这篇文章.)斯蒂阶从未发表他的证明的事说明他最后发现了自己的错误.但不管怎样,1897年F·默顿斯(F. Mertens)发表了50页的 n 直到10 000的 $\mu(n)$ 和 $M(n)$ 值的表,并得出结论:不等式(9)确实是"非常可能"的.所以这个猜想现在叫做默顿斯猜想.

在1897年到1913年的一系列论文中,冯·斯特恩耐克(von Sterneck)发表了 n 直至五百万的 $M(n)$ 的值,并发现这些数据全都满足默顿斯猜想.他认为当 n 超过200以后,就满足更强的不等式

$$|M(n)|<\frac{1}{2}\sqrt{n}.$$

但在1960年,卓卡特(Jurkat)证明了这是错误的:使 $|M(n)|\geqslant\frac{1}{2}\sqrt{n}$ 的最小 n 值是 $n=7\ 725\ 038\ 629$,此时 $M(n)=43\ 947$.

接着在1979年,科恩(Cohen)和德来斯(Dress)计算了 n 到78亿的 $M(n)$ 的值,并注意到所有的值都满足不等式

$$|M(n)|<0.6\sqrt{n},$$

这似乎又说明默顿斯猜想可能是真的(公平地说,应该提到用其他方法也得到了大量的证例).但1983年10月赫耳曼·特·黎勒和安德鲁·奥德莱斯科(Andrew Odlyzko)用8年的合作成功地证明了事实不是这样.

他们的结果是由古典数学方法和能力强大的计算机相结合取得的.不仅如此,他们的合作是今日科研的一种日渐明显的趋势的范 [218]

例.在大多数时间里,他们都在自己的工作基地(黎勒在阿姆斯特丹数学中心,奥德莱斯科在新泽西贝尔实验室)通过电子信函联系.

他们是怎样得到结果的呢?当然不是找到了一个使$|M(n)|\geq\sqrt{n}$的数n.在当时,这样的数还没有发现.能得到的例证表明:在10^{30}以下没有这样的数.他们证明的关键是1942年因格汉姆(A.E.Ingham)得到的结果.

首先,注意到不等式$|M(n)|<\sqrt{n}$可改写成

$$\frac{|M(n)|}{\sqrt{n}}<1.$$

因格汉姆所作的是说明如何定义一个函数$h(x)$,使其有下面性质:对任意实数x,比$h(x)$小的数不能大于每个$|M(n)|/\sqrt{n}$的值.所以为了否定默顿斯猜想,只需找到一个x使$h(x)>1$.假设你找到一个x使$h(x)=1.06$,则因格汉姆的结果说明所有比1.06小的数都不大于任何一个$|M(n)|/\sqrt{n}$的值.特别地,1(小于1.06)不大于任何$|M(n)|/\sqrt{n}$的值,这与默顿斯猜想矛盾(但注意这里没有给出一个使默顿斯不等式失效的n值).

函数$h(x)$的定义涉及到一些非常抽象的数学,包括黎曼ζ函数.特别地,对给定值x,$h(x)$的计算牵涉到ζ函数零点个数的相当精确的计算.(因格汉姆有关$h(x)$函数的结果要求所有被计算过的零点都符合黎曼假设——虽然如果期望这个假设正确的话可能不会担心这方面的事,不管哪种情况这个假设已在超过了为否定默顿斯猜想所需要的区间上被检查过.)无疑,这种计算要是没有强大的计算机是太耗费时间了.即便如此,在写好一个对任给的x完成$h(x)$的计算程序的同时,还必须面对找个使$h(x)$比1大的x的任务.这[219]个部分是更加困难的.函数$h(x)$表现出的趋势令人失望:得到的值几乎总是远远小于1.事实上,到1979年,黎勒能得到的最好的值是0.86.看来找一个合适的x就像大海捞针一样困难.在那时,黎勒下结论说这个问题超出了现存计算机的计算能力.

事实说明,这个问题的突破并非由于新的计算机技术,而是伦斯

特拉(Lenstra)和拉瓦茨(Lovász)1981年发现的一个有力的新算法(它被公开后得到了广泛的应用).在应用到默顿斯猜想上时,这项新技术(这里不去描述)正好提供了为寻找合适的 x 值所需的计算手段.黎勒和奥德莱斯科从此有了大海捞针的法宝.

证明的第一步是计算出一些 ζ 函数的零点.这些零点是计算他们可能会在搜索中碰到的 $h(x)$的值所需要的.为此一台阿姆斯特丹数学中心的 A CDC CYBER 750 型计算机工作了40个小时,得到了精确到100位数字的2000个零点.接着,应用上面提到的新算法,贝尔实验室的一台 CRAY-1 计算机工作了10个小时,直到发现了一个使 $h(x)$比1大的 x 值.它使 $h(x)=1.061545$.默顿斯猜想最终被否定了.据记载,达到目的的 x 值是个在小数点前有65位数字的大数:

-14 045 289 680 592 998 046 790 361 630 399 781 127 400 591 999 789 738 039 965 960 762.521 505

当然,这个结果不仅否定了默顿斯猜想,它还表明不等式

$$|M(n)|<1.06\sqrt{n}$$

不总成立.那么有没有另一些常数 A 使不等式

$$|M(n)|<A\sqrt{n}$$

总成立呢?如果有一个 A 值使不等式成立的话(对所有 n),则黎曼假设成立.为了用黎勒和奥德莱斯科所采取的方法证明这种形式的 [220] 一个不等式错误,必须找到一个使 $h(x)$大于 A 的 x 值.黎勒和奥德莱斯科都相信,不管 A 有多大,这在理论上都是可行的——就是说,他们相信 $h(x)$可以达到任意大的值.不过他们提醒说,用当今最新的算法和计算机技术,在实际中所能希望的最好的 x 值,也只能让 $h(x)$达到1.5左右.他们还肯定说,达到2已是超过我们现实能力的事情.

那么随着默顿斯猜想的否证,黎曼假设的情况如何呢?完全依然如故.如果默顿斯猜想正确(如斯蒂阶所认为的那样),黎曼猜想的正确性就唾手可得.但知道默顿斯猜想不成立,却不能对黎曼

假设作任何结论．这两个猜想：默顿斯猜想和黎曼猜想，它们是不等价的．

然而默顿斯猜想有一个减弱条件的叙述，它等同于黎曼假设：对任何大于1/2的实数r，存在一个常数A使不等式

$$|M(n)| < An^r$$

对所有n都成立．（换句话说，把默顿斯猜想中的指数换成大于1/2，就会得到一个相似类型的不等式．）这一命题如有错误当然也就暗示了黎曼猜想的错误（就像它的正确暗示黎曼猜想的正确一样）．但这是与默顿斯猜想截然不同的问题．

比贝巴赫猜想

数学，像它通常呈现在世界面前的，是一种冰冷的，没有人情味的知识．数学真理的独特之处在于它不因时间、地点和人物的改变而改变．但是，数学的发展过程却是一种人类的活动，所以它受到各种影响的制约．虽然数学真理的绝对性最终是不能被否认的，但有时为[221]达到最终的目标却要花费一些时间．

假如你是一位世界数学界的领导人物．你在某一问题上已钻研了多年，有好几次你已经近乎解决了这个问题．一天，你收到一份385页的打字手稿，主要意思是解决了你那个问题．你看看上面的署名，啊，一位令人尊敬的数学家．这可不是你每天都能收到一打，然后马上扔进废纸篓的那种思想怪诞的文章．不过这个人已有52岁了，而传统上认为（根据一些颇令人佩服的统计资料）数学家一过了40岁左右就不会有什么太出名的工作了．还有，这位数学家以前也曾声称自己解决了某某问题，而有好几回最后在讨论中发现了严重的错误．而现在他又声称解决了那个在70年的历史中不仅难倒了你，这个世界上这一问题方面的著名专家，而且还击败了包括世界上许多其他数学界领导人物的问题（有几个人也曾一时以为他们找到了答案）．草草看过一遍手稿后，你知道你的这位同事用了你和熟悉这一

问题的所有人都认为最不可能成功的一种繁琐的方法.

面对这样的情况——无疑现在你还有很多别的事要做——你该怎么办?这就是美国数学家卡尔·菲茨杰拉德(Carl FitzGerald)在1984年春天所处的位置.他的同胞路易·德·布朗日(Louis de Branges)真的像他宣称的那样,已经解决了比贝巴赫(Bieberbach)猜想,这个许多人都没有解决的大问题了吗?菲茨杰拉德觉得好像不太可能.("我不希望他的证明是正确的,"他后来写道,"我有两篇文章说明比贝巴赫猜想至少接近于正确……我不想让我的结果被取代.")德·布朗日还给全国的另外十来个数学家寄去了手稿副本,可是他们也都抱着同样的怀疑态度.

但是事有凑巧,作为苏美交流计划的一部分,德·布朗日被优先安排访问苏联.在访问期间(同年4月到6月),他被安排到列宁格勒大学作报告.这是阐述他宣称的证明的理想场所.在他的听众当中将有在这一问题上的三位世界著名专家米林(I.M.Milin),埃梅拉诺夫(E.G.Emel'anov)和库茨米那(G.V.Kuz'mina).1971年米林本人曾提出过一个猜想,他还阐明了比贝巴赫猜想是它的一个推论.而德·布朗日所证明的实际上就是米林所提出的那个猜想.俄罗斯数学家虽然也是充满怀疑,但却是耐心的听众,他们听完了五次每次四个小时的系列报告.他们期望着会随时挑出错误. [222]

可是发现错误的一刻没有到来,因为没有错误,证明是正确的.下一步是看能否把冗长的证明简化.经过研究,他们进行了一些出色的修改,把文章的长度缩短到仅有13页.这13页的论文向世界各地寄发后,人们经历了一个坐直——注意——相信的过程.德·布朗日所做的是许多数学家可望而不可及的工作.他解决了一个长期悬而未决的、公认的"难"题.

事实上,这一问题的难度是它所以闻名的主要原因,就目前所知,比贝巴赫猜想并不像黎曼假设那样会导致许多精采的进展.但它是一个"干净"的结果,显示了数学及复分析的有序性.它的内容叙述如下:

设有无穷级数

$$B = x + a_2x^2 + a_3x^3 + \cdots,$$

x 是复变量，系数 $a_2, a_3, \cdots$ 是复数（因 x 系数是 1，$a_1 = 1$）. 对变量 x 的任一给定值，无穷级数能够产生一确定的答数（即级数的"和"）. 但为此要求级数的项以很快的速度变小——以抵消项数是无穷的这种作用.

例如，以前给出的 $\sin x$ 是 B 型的级数：

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$(a_2 = 0, a_3 = 1/3!, a_4 = 0, a_5 = 1/5!, a_6 = 0, \cdots)$. 对 x 的任意值，该无穷级数给出有限的答数，因为系数 a_n 以如此速度变小.（同样，e^x 的级数虽然第一项是 1，不是 B 型的，但如减去首项的 1，得到 $e^x - 1$ [223] 的级数就成了 B 形的. 像 $\sin x$ 一样，对任意 x 值给出有限和.）

但是，如果数列中的系数不以这样快的方式减小，那么无限和的存在就取决于取作 x 值的数的大小了. x 值取得小，可能会得到有限的答案，x 值取得大，会得到无限的结果（有些情况还根本没有结果）. 特别地，若 x 小于 1，则当 n 增大时 x^n 减小，就有可能得到一个有限的结果. 这完全取决 a_n 的系数. 所以，数学家研究 B 形的数列时，他们经常把注意力放在使 n 增大时 x^n 变小的 x 值的数列的表现上，即他们只考虑绝对值小于 1 的 x.（复数 x 的绝对值是指它在复平面上到原点的距离，用 $|x|$ 表示. 这一符号原用来表示把实数 x 去掉负号. 其实想想这两个用法并不矛盾. 复数的新用法只是扩展了旧的用法. 直接在一直角三角形中应用毕达哥拉斯定理，如果 $x = a + bi$，则 $|x| = \sqrt{a^2 + b^2}$.）

在复平面上，所有使 $|x| < 1$ 的复数 x 的集合形成了一个半径为 1 的圆盘，中心位于原点，通常叫做单位圆盘.（因此说 x 位于单位圆盘内只不过是 $|x| < 1$ 的几何表述.）

假如现有一 B 形级数，而且级数对单位圆盘内的每个 x 值都有一有限的（复数）和（以 $f(x)$ 表示），这级数便定义了一个函数 $f(x)$，

使对单位圆盘上的每个 x 都指定了一个 $f(x)$ 的值. 函数 f 作用在不同的 x 值上时可能得到同一结果. (比如设级数 B 有 $a_2=1$, 而其他系数为0, 即 $f(x)=x+x^2$, 则检验可知 $f(-1/3)$ 和 $f(-2/3)$ 都等于 $-2/9$.) 如果不同的 x 值总给出不同的 $f(x)$, 则函数 $f(x)$ 叫做一一的(基于这样的事实, 即 $f(x)$ 的每一个值是从唯一的一个 x 值得来). 这一特性非常有用, 所以数学家给它这样一个特殊的名字(有点像多数现代社会中的一夫一妻制).

现在假设级数 B (不管它实际是什么)产生一个一一的函数 $f(x)$, 使对单位圆盘上的任意 x 有有限的值, 就能在复平面上画出 $f(x)$ 的几何图象. $f(x)$ 把单位圆上的每一点 x 与复平面上的某一点 $f(x)$ 联系起来. 这样得到的 $f(x)$ 的集合是复平面上的一个区域(子 [224] 集). 那么我们很自然地要问这个区域有多大. 因为 x 值取自单位圆盘, 而单位圆盘比整个复平面小(后者在各个方向伸向无穷). 你也许会想 $f(x)$ 的值域不过是整个平面的一个小部分罢了. 可是别忘了, 单位圆盘与整个复平面都包含着无穷多的点, 就像在第2章充分展示的, 无穷集合的行为与我们熟悉的有限集合完全不同. 这里很可能也不一样. 上面问题的答案是: $f(x)$ 的值域实际上有可能是整个复平面! 根本无需特别构造函数 $f(x)$, 现成的柯布(Koebe)函数就能做到这一点, 它可以通过以下公式计算:

$$K(x)=\frac{x}{(1-x)^2}$$

或由以下的 B 级数得到:

$$K(x)=x+2x^2+3x^3+4x^4+5x^5+\cdots.$$

(因柯布函数源自 B 形的无穷级数, 是一一的, 它正是我们观察过的那种函数.) 柯布函数给出的与单位圆盘上的 x 对应的所有 $f(x)$ 值的集合是由除去实轴上的 ∞ 到 $-\frac{1}{4}$ 的点以外的所有复数组成的. 所以在几何上, 这一函数要多大就有多大(这里"大"指值域的大小).

柯布函数是值域最大的"纪录保持者", 那么是不是它的级数中

每一项的系数也是最大的呢？也就是说，取一 B 形级数，它定义了一个单位圆盘上的所有 x 的(有限的)单值函数，是否有：

$$|a_2| \leqslant 2, |a_3| \leqslant 3, |a_4| \leqslant 4, \cdots?$$

(或者把问题换个提法，如果级数 B 有一个或多个系数不能满足相[225]应的不等式．比如 $|a_{163}| = 163.5$ 则是否会得出：要么 f 不是单值的，要么更糟，单位圆盘上有些 x 使 B 级数没有有限值?)

这就是本世纪初德国数学家路德维希·比贝巴赫(Ludwig Bieberbach)提出的问题．在 1916 年的一篇文章中他猜想答案为"是"．但是他只能证明他的猜想的第一个系数．即他设法证明了 $|a_2| \leqslant 2$.

英雄故事就此开始．这一猜想看起来相当有道理，是一种复分析经常产生的"干净"的结果．它的陈述十分简明，保证有不少数学家跃跃欲试，但发现最终的证明却花了很长时间．

在比贝巴赫之后，有关这一猜想的第一个结果是德国人查尔斯·吕维内尔(Charles Löwner)作出的，1923年他找到一种方法证明了系数 a_3时的情况(即他证明了 $|a_3| \leqslant 3$)．1955年，在美国工作的哥拉比但(Garabedian)和希费尔(Schiffer)证明了下一个系数 a_4时的猜想．1968年，佩得森(Pederson)和奥兹华(Ozawa)跳过了 a_5证明了 a_6时的情况．1972年佩得森和希费尔合作处理了略过的 a_5．同年奥兹华和库博塔(Kubota)丢下 a_7证明了 $|a_8| \leqslant 8$，这标志着步步逼进的结束．

这的确是缓慢的进展．如果像这样一个一个地检验每个系数，那么整个猜想是不可能被证明的(因为系数是无穷多的)．但是这种进展最后总是能碰巧引出解决这一问题的方法．当路易·德·布朗日在 1984 年最终证明了比贝巴赫猜想时，他的方法实际上是以一些早期的工作为基础．早到什么时候？——1923 年吕维内尔的工作，比贝巴赫本人之后研究这一问题的第一人！现在我们要在故事的主要情节之外加一段插曲，因为当攻克那些系数的过程在一个一个地缓慢进行之时，另一个过程也在悄然进行．

这个想法是探讨下列形式的不等式：

$$|a_n| \leqslant Cn.$$

这里 C 是一些常数，要看看对所有的值，不等式被验证都成立时，C 能达到多小. 为证明比贝巴赫猜想，需要说明 C 有可能是 1. 可是在此之前，我们能与这个值接近到什么程度呢？ [226]

J·E·利特伍德发现了第一个这样的结果. 1925 年他证明了 $C = e$ 是可能的（e 约等于 2.718）. 从那以后这一值被逐渐降低. 1956 年，前面提到的米林说明了 C 可以为 1.243. 1972 年，菲茨杰拉德（在本节的开始，正是他接到了 385 页的手稿）把 C 降到 1.081. 1978 年，他的学生戴维·赫罗维茨（David Horowitz）又降至 1.066. 已经到了非常诱人的接近程度，但还近得不够. 1984 年比贝巴赫猜想最后完全解决时，却不是靠这种"减小 C"的过程，而是根据 1923 年吕维内尔原来的工作.

为了证明 a_3 的情况，吕维内尔引入了一个偏微分方程，它的解可以逼近单位圆盘上的任一单值函数. 现在的做法是把数学翻译成物理语言，如河中的水流，微分方程表示了一个扩张流，有可能沿流"传递信息"——特别地，能估计这个函数的 B 级数的系数的大小. 德·布朗日在证明中引入了一些辅助函数 $t_1, t_2, \cdots$ 以掌握所要的信息，每个 t_k 由涉及所有前面到 t_{k-1} 的函数的微分方程所定义. 比贝巴赫猜想的证明（实际是前面提到的较强的米林猜想的证明）则化约为对这些 t 函数满足某些条件的证明. 这是对这一问题的极重要的、决定性的，然而又是老式的证明方法，就好像是一场精采的马戏，德·布朗日竭尽全力要使所有的球保持在空气中一样，但它却成功了.

阅读文献

复数导论的内容可以在许多数学基础读物中找到，比如 Keith Devlin 的 Sets, Functions and Logic（Chapman & Hall, 1981）. 同样也有复分析的高级一点的知识的著作，如 G.J.O.Jameson 的 A First Course on Complex Functions（Chapman & Hall, 1970）.

K. Chandrasekharan 的 Introduction to Analytic Number Theory (Springer – Verlag, 1968) 一书如其书名所示，虽然书中的实质性内容并不是外行们所能搞清

[227] 的.

有关黎曼 ζ 函数的著作是 H. M. Edwards 的 Riemann's zeta function (Academic Press, 1974).

A. M. Odlyzko 和 H. J. J. te Riele 的研究文章 Disproof of the Mertens conjecture 发表于 Journal für die reine und angewandte Mathematik, Volume 357 (1985), pp. 138 – 60. 描述了默顿斯猜想的解决过程.

比贝巴赫猜想及其解决的简要叙述可见 O. M. Fomenko 与 G. V. Kuz'mina 的文章 The last 100 days of the Bieberbach conjecture (The Mathematical Intelligencer, Volume 8 (1986), pp. 40 – 7). 还有个记载见于菲茨杰拉德的文章 The Bieberbach conjecture: Retrospective (Notice of the American Mathematical Society,

[228] Volume 32 (1985), pp. 2 – 6.)

（李家宏译）

第10章 纽结与其他拓扑问题

童子军,物理学家和其他

怎样识别平结与错平结?一个普通的童子军可以毫无困难地回答这个问题.但数学家能进行这样的区分吗?1984年关于这问题的研究取得了重要突破.

物理学家在研究我们生活的四维时-空宇宙时使用的数学是否正确?1982年以前,每个数学家都会异口同声地回答:"当然正确,这样的数学只有一种."然而现在我们了解更多了.还有其他不同类型的数学可以被应用于四维宇宙——但仅仅是四维宇宙;对二维空间或三维、五维、六维以及更高维空间来说决不会发生这种情况.四维空间的特殊性不仅在于我们的宇宙似乎是四维的,而且还在于数学上它也很特殊——不过表现形式完全出人意料.

这些只是近年来在称作为"拓扑学"的数学领域中所发生的两项进展.这门学科范围如此广泛,介绍它可能需要整整一本书而不是眼前这一章,书名可叫《拓扑学的新问题——新的黄金时代》.今天,拓扑学,至少是它的某些分支,正在向数学的大多数领域渗透,更不用说它与现代物理学的许多深刻联系了.这门学科在本书第7章讨论四色定理时已经亮相.(与其他各章专题的联系尚未提及,但这种联系确实存在.)这门学科只有一个世纪的历史.虽然有些思想可以追溯到欧拉与高斯,但其真正的发展道路是19世纪晚期由H·庞加莱(Henri Poincaré)和其他一些人的工作而开拓的. [229]

本章只涉及拓扑学的两个方面：纽结理论——一个引人入胜但却高度专门化的领域，以及流形理论——关于几何曲面和广义曲面的性质的研究．大多数拓扑学家会把流形理论看作是他们这门学科的核心．（作为应用流形理论的突变论，是本世纪60年代由拓扑学家R·托姆（René Thom）和C·齐曼（Christopher Zeeman）创立的，虽然它完全符合本书的选材标准，但还是被略而不讲，我的理由是关于这一专题的优秀通俗读物很容易找到．①）

虽然拓扑学是一个极为深奥的研究领域，但对有兴趣的门外汉（也许就是读者您本人）来说，只要有几何对象的直观能力，就能理解其一般原理．因此以下的介绍采取了高度几何化的手法．

什么是拓扑学？

拓扑学是数学中一个很容易定义又很难理解的分支．拓扑学可以简单地定义为研究几何对象在连续变换下保持不变的性质的学问．直观地说，所谓连续变换（也叫拓扑变换）就是使几何对象受到弯曲、拉伸、压缩、扭转或它们的任意组合．这里理想地假设受到变形的几何对象具有弹性，从而能充分经受这样的操作．必须附加的条件是：在变换前连在一起的点变换后仍应连在一起．经过适当的处理（这并不容易），上述要求不像乍看那样与几何对象可任意拉伸这一点相矛盾．需要避免的是对象的任意切割，撕分或“粘连”．至少，在一般的意义上是这样：将对象切割后再沿切口粘连起来，这样的变换是允许的，以使原先连在一起的点仍然连在一起，否则有些操作就不可能进行．例如，为了把图48(i)所示的形体连续变换为图48(ii)所示的形体，允许将两个套在一起的环像图48(iii)那样切割开来，然后再将两边的切口重新粘连在一起．两个可以这样互相变换的形体就叫做是拓扑等价的．（事实上就上述特殊的例子而言，切割并无必要．

[230]

① 如《突变论》，A．Woodcock 和 M．Davis 著（Penguin 1980）．——原注．

仅仅通过拉伸与弯曲就可将图48(i)变换成图48(ii).你知道怎样做吗?答案如本章结尾图62所示.这个例子说明了即使像“简单拉伸”这样的概念也是难以捉摸的.) [231]

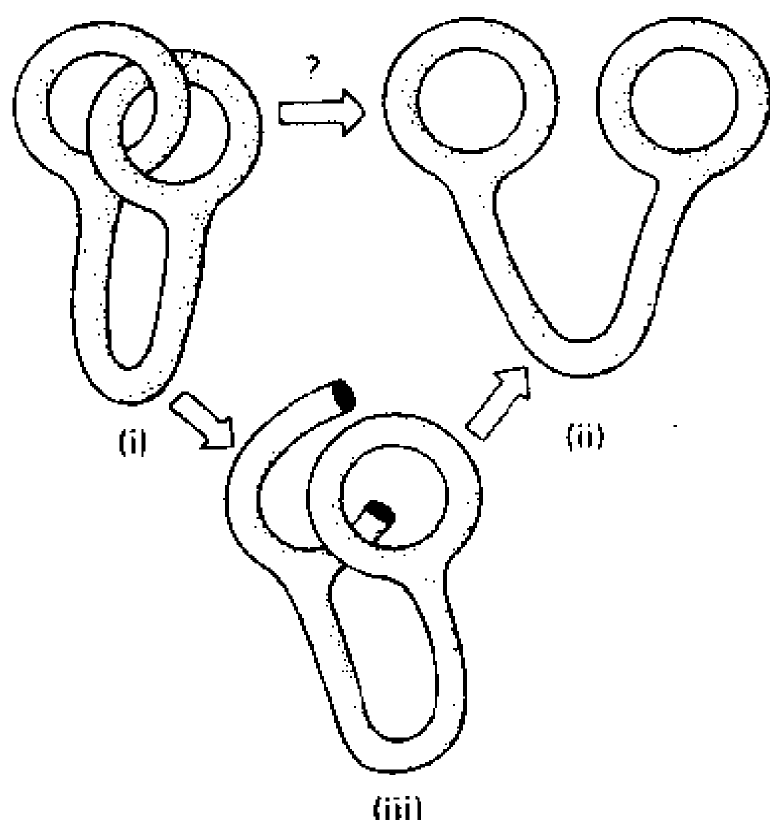


图48　环的难题.设有如图(i)所示用高度弹性的物质做成的物体,你能通过变形而使两个环像图(ii)那样解开吗?最显而易见的方法是将其中一个环切开,使两环分离,然后再将两端的切口重新粘连起来.只要两自由端的粘连方式与分割前完全一样(即任何一端都不受扭转),上述过程就是将(i)变成(ii)的一种拓扑变换.但实际上不用切割就可能使(i)变换成(ii),只需以适当的方法对物体作出处理即可.你晓得怎样做吗?图62给出了一个解法.

“拉伸-切割-粘连”这类拓扑描述之所以不够恰当,原因之一是它只能被应用于曲面(即二维形体)的研究,而拓扑学却要处理任意维数的形体,包括三维、四维、五维和更高维的情形.即使对于曲面

来说也有困难.因为习惯于在三维空间中来考虑二维形体,这就容易将二维形体的性质与背景空间的性质混淆起来.确实,维数概念本身就常常引起混淆.例如在本章中所谓球是指二维的球面而不是球体.这种曲面上的点只能沿该曲面上的两个独立方向移动.然而,一个球面的具体作图却只能在至少是三维的空间中才能实现.

[232] 还可以举另一个例子:图 49 中的球面与环面①在拓扑上似乎是有区别的(它们也确实不同).任何拉伸、弯曲或切割、粘连似乎都不可能将其中一个变换成另一个.(请记住唯一许可的切割－粘连过程要求所有被切开的点必须重新粘连起来,因此决不可能通过先切断

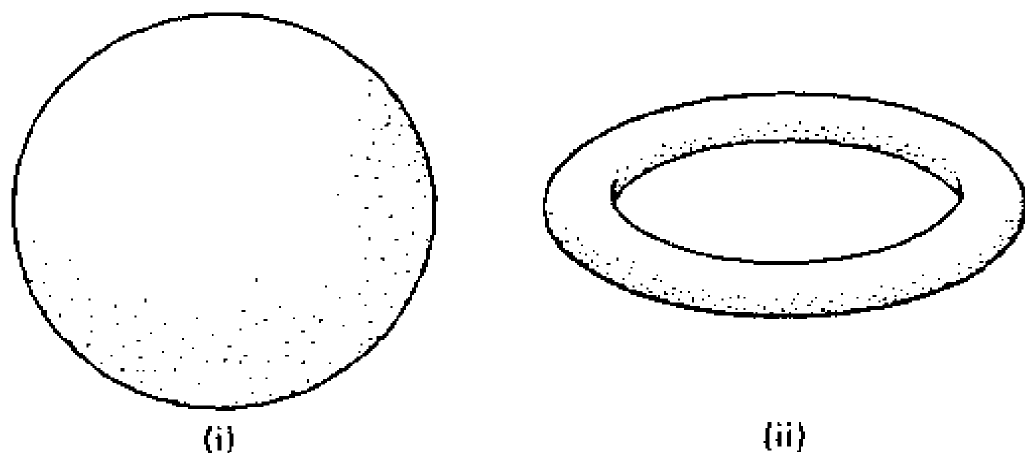


图 49 球面(i)与环面(ii),在拓扑学中二者皆被看作"中空"曲面而不是实心立体.大多数人会同意这样的看法,即不可能用任何拓扑方法将其中一个变为另一个——也就是说这两个曲面在拓扑上是不等价的.为了证明这一点,你必须指出其中一个曲面的一种拓扑性质为另一个曲面所没有.环面的"洞"似乎属于这类性质.但问题在于这个洞不是环面而是背景空间的性质.区分这两种曲面拓扑性质的是所谓的欧拉示性数.虽然环面的欧拉示性数的确是曲面自身的性质,但它如实反映了环面(不如说不是环面)的洞.

---

① 像球面一样,环面也被看作是中空的曲面.——原注.

再将两端粘连起来而将一个环面变换成球面.)这两种曲面的基本区别似乎是在于:环面有一个"洞"而球面则没有.是不是这样呢? 否.环面是一种光滑的曲面,上面根本不存在"洞".如果你被限制生活在一个巨大的环面上,你可以走遍整个曲面而决不会遇到任何洞.上面提到的洞是与这类特殊曲面落在三维空间中的方式相关的.换言之,这个洞是某种属于背景空间的东西.这并非说这个洞与环面的拓扑毫不相干,只是说它不是环面自身的性质.环面有一种拓扑性质与背景空间中洞的存在密切相关,下面我们就来介绍这种性质(它被用作为区别环面与球面的特征).

曲面自身的性质与其背景空间的性质之间的区别是很微妙的.这一点若与"边界"这个拓扑概念相比较就很清楚.无论是球面还是环面都没有边界,它们都是所谓闭(无边界)曲面.一个圆盘有一条边界,带一个洞的圆盘(像一张唱片)有两条边界.如果取一条纸带将其两端粘连在一起形成圆柱带(如图50(i)),就可得到一个有两条边的曲面.如果将纸带拧半转后再将两端粘连起来(如图50(ii)),可以得

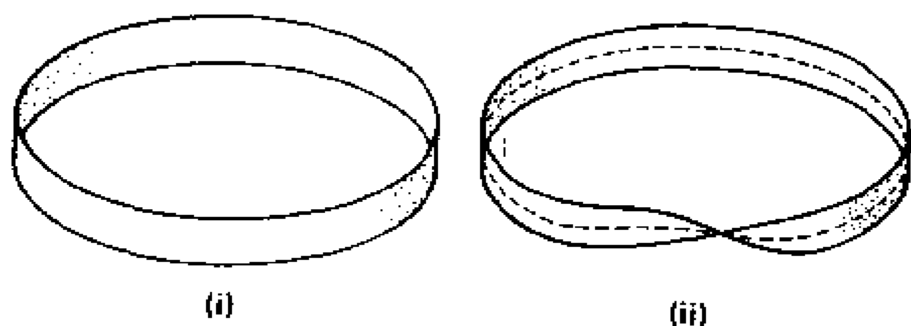


图50 圆柱带(i)和莫比乌斯带(ii).圆柱带是具有两条边界的可定向曲面的例子;而莫比乌斯带则是单边、不可定向曲面的例子.要从一条纸带做一个莫比乌斯带,只需将此纸带扭半转,然后再将两端粘连起来.莫比乌斯带有一些奇特的性质.例如你可以看一看当莫比乌斯带沿上图所示虚线剖开后会发生什么情况.结果将完全出乎你的意料.你还可以类似地去剖分另一个莫比乌斯带,但这次是沿离边缘三分之一宽的虚线剖分.

到一个只有一条边界的曲面（试做一个瞧瞧），叫莫比乌斯带，是以第九章中已提到过的数学家莫比乌斯（A. F. Möbius）的名字命名的.

将两条莫比乌斯带边对边地粘连起来，结果得到的就是所谓的克莱茵瓶（见第7章图44）. 不过在三维空间中不可能进行这样的作图，除非允许曲面能够穿过自身. 无独有偶："莫比乌斯带（和克莱茵瓶）是只有一个侧面的曲面."用这种说法来表达曲面拓扑的基本概念也颇容易引起误解，因为"侧面"的概念过分依赖于物体落在三维或更高维空间里这一认识（从空间的高处可以俯视曲面的侧面）. 一个莫比乌斯带或克莱茵瓶乍看似乎有两面，然而你会发现要给这两 [233] 面涂上不同的颜色是不可能的（至少请对莫比乌斯带试一试，看看会发生什么情况），这当然会有助于认识这些曲面的非同寻常的性质，但却仍不能说明曲面真正的拓扑特征. 在目前的情形，这种特征并不是"侧面性"（sidedness），而是"可定向性"（orientability）. 所以从现在起请尽量不要再把曲面想象为有"侧面"的东西.

一个曲面如果可以在下述意义上将"顺时针"与"逆时针"的概念区分开来，就被称作为可定向的. 设在曲面上（最好是曲面内，务请记住曲面无"面"！）画一小圆并给定一个方向，如图51所示. 那么无论你怎样在曲面上（最好说曲面内）移动这个小圆，都不能使它的方向变反.（形容词"小"的意思是指圆充分小以致能沿曲面自由移动而不会遇到诸如洞或隆起之类的东西.）圆柱带（图50(i)）是可定向的；莫比乌斯带（图50(ii)）是不可定向的. 为了明白这一点，应用某种透明材料（如市售供投影仪使用的"醋酸胶片"就很合适），以便从曲面的两"侧"皆可看见你所画的定向圆，因而可以认为它们是在曲面"内". 现从画在带上的一个带箭头的小圆出发，按一定间隔相继复制这个小圆（包括箭头），这样在带上"移动"小圆. 对莫比乌斯带来说（见 [234] 图52），当你回到出发之处（你可能会误认为是在其反面），将会发现小圆上箭头的方向已经改变：上述一连串的圆表明了"顺时针"方向是怎样在小圆从未离开过曲面的情况下变成"逆时针"方向的. 这就

是所谓不可定向性.对于圆柱带来说,就不可能发生这种反向,而这就是可定向性.

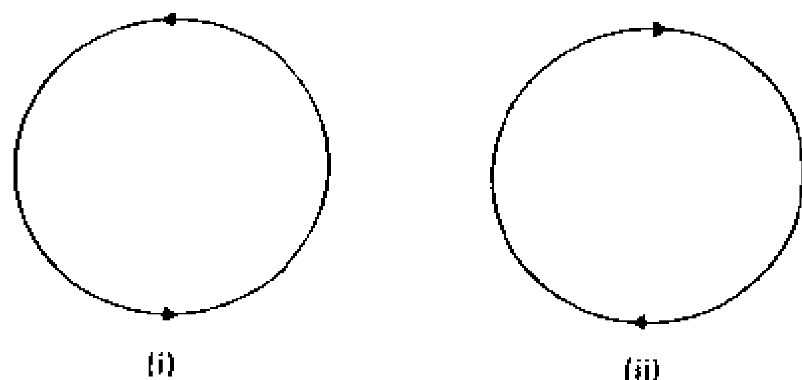


图51 可定向性.一个曲面被称为可定向的,如果不可能通过在曲面上的移动而将逆时针指向的圆(i)变成顺时针指向的圆(ii).可定向性是曲面的一种拓扑性质,它对应于"双面性"的直觉概念(如圆柱带的情形);而不可定向性则对应于"单面性"概念(如莫比乌斯带的情形).

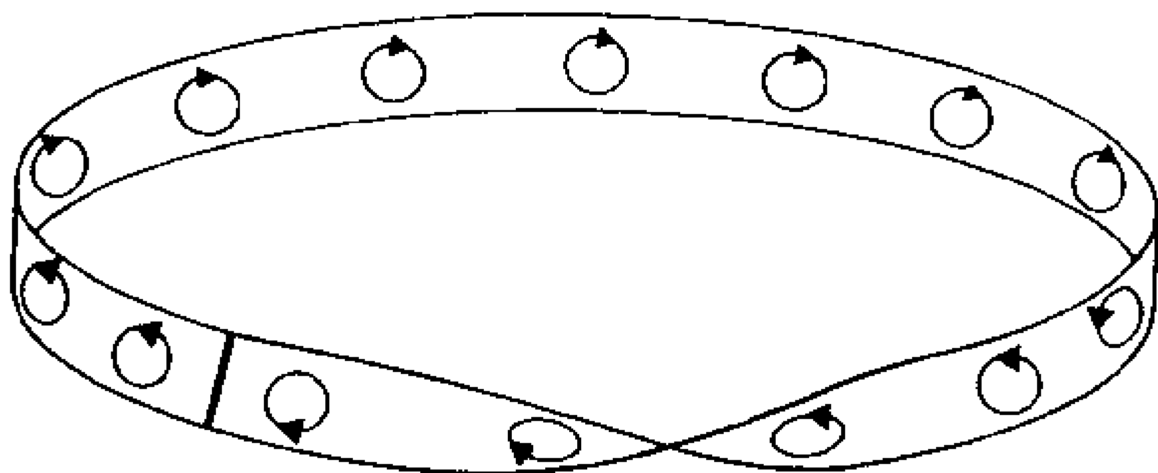


图52 莫比乌斯带的不可定向性.用一条透明材料的带制作一个莫比乌斯带,看一看当一定向圆在带上移动时将发生什么情况.

235

怎样研究拓扑学?

对这问题的一个滑稽但却十分中肯的回答是:"多加小心".拓扑学允许非常一般类型的变换,这意味着几何图形大多数众所周知的

性质将不再适用.古典几何(可以被看作是研究图形在刚体变换——平移、旋转与反射下保持不变的性质的学问)中常用的概念包括直线、圆、角、长度、面积和垂直等等,所有这一切在拓扑学中都失去了意义,因为它们在连续变换下可以变得面目全非.

在古典几何中,为了证实两个几何图形是否相同,要看你能不能(在理论上)通过刚体变换来移动其中一个图形,使它与另一个图形在所占位置上重合.如果能找到这样的变换,就说这两个图形"相等"或"全等",或者更正式地说它们是几何"等价"的.为了说明两个图形不相同,则需找出某种特殊的几何性质,它对于两个图形来说有所不同,例如一个图形比另一个图形大,或二者某个角度不相等.当然,所有这些都已人所共知,似乎不值得一提.我们一直是用这样的方法来进行图形比较的.然而,拓扑学中所采用的正是完全同样的方法,只不过用到的变换的概念有所不同(并且对我们来说很不熟悉)罢了.

为了证明两个图形拓扑等价[①],需要找到一个拓扑(即连续)变换,使其中一个图形变成另一个.例如三角形与圆是拓扑等价的,它们之间存在着显见的变换方式.为了证明两个图形在拓扑上不同(即不等价),则需找出某种只为其中一个图形所独有的拓扑性质.例如边缘,圆盘有一条边缘,球面却没有边缘,因此二者在拓扑上是不同的.(你不可能通过连续变形将圆盘变成球面,反过来也如此.)为了[236]在拓扑上对两个闭曲面(即没有边缘的曲面)如球面与环面作出区别,则需要找出其他某种拓扑性质.环面的"洞"不能起这样的作用,因为正如我们已经看到的那样,它并非曲面本身的拓扑性质.可定向性同样也无济于事,因为这两种曲面都是可定向的.(由于克莱茵瓶是不可定向曲面,可定向性这一概念可以被用来将球面、环面与克莱茵瓶区别开来.)

因此,拓扑学的首要任务之一就是要找出足够的拓扑性质,使我

① 数学家们使用术语"相等"和"全等",但对于初学者来说,它们带有太强的几何内涵,这里尽可能地避免使用.——原注.

们能对任意两个不等价的图形作出区别,方法是证明只有其中的一个图形具有某种特别的性质.当然,任何这种性质对于所有与被考虑的对象拓扑等价的图形来说必须相同.也就是说拓扑性质必须在连续变换下保持不变.(如有改变,就不是拓扑性质!)正因为如此,这些性质通常就叫拓扑不变量,用这名称比叫"性质"好,因为许多不变性质都是数值不变量.可定向性是一种拓扑不变量;曲面的边缘数也是拓扑不变量.(有一条边缘的曲面不可能与有三条边缘的曲面等价.)还有一种拓扑不变量,它与上述两种不变量一起,足以被用来区分所有的非等价曲面.但在讨论这种不变量以前,有必要对像"球面"(the sphere)或"克莱茵瓶"(the Klein bottle)这样(带定冠词)的说法作一解释.因为(例如)任意两个球面拓扑等价(它们只是位置和大小有所不同,而后二者都不是拓扑不变量).从拓扑学家的观点看这两个球面是"全等"的,因此说"the sphere"是有意义的,即使你所观察的图形像一段香肠.当然,如果同时考虑两个球面,加定冠词就不合适,但若你的目的只是简单地表达球面这一概念,那么"the sphere"这种说法就是有意义的.

现在来谈谈曲面的第三种不变量.事实上,我们在第7章中已经遇到并使用过这类不变量.正如在第7章中所证明的那样,对任意曲面来说,由覆盖(整个)曲面的一张地图或网络的顶点数(V)、边缘数(E)和面数(F)所得到的数量 $V-E+F$,与具体网络的选取及它在曲面上的位置无关.同时可以证明(并且也容易使人相信),此数在曲面的任何连续变换下保持不变,因此是曲面的一种拓扑不变量,叫曲面的欧拉示性数.

球面的欧拉示性数是2(即对与球面拓扑等价的任意曲面和覆盖曲面的任意网络来说,$V-E+F=2$).环面的欧拉示性数是0,因此环面与球面在拓扑上是有区别的.与环面的"洞"有关的恰恰是欧拉示性数.双环面有两个洞,其欧拉示性数为 -2,三环面有三个洞,其欧拉示性数为 -4,等等.一般公式将在后面给出.克莱茵瓶的欧拉 [237]

示性数与环面①一样是0，因此，该不变量不能用以区分这两种曲面(但正如已经提到的那样，在这一情形可定向性可担当这方面的角色)．三种不变量——边缘数、可定向性和欧拉示性数在一起，足以对所有的(二维)曲面进行区分．

纽结的拓扑

有些拓扑学家可能认为纽结拓扑不是“拓扑学”．在某种意义上他们的看法是对的．纽结理论是拓扑学，但却是一种非常特殊的拓扑学．纽结理论研究的对象(纽结)必须是三维空间中的形体．在二维空间中不可能做这样的纽结——那里没有充分的“余地”使你能将一根绳子或其他任何类似物自身缠结．在四维或更高维空间中则“余地”太多，以致任何一个纽结都会被立即打开而变成没有纽结的绳子．另外，对纽结施行的拓扑变形还必须特别排除一般拓扑变换所允许的切割和粘连(作这种限制的原因不言而喻)．

那么究竟什么是纽结理论呢？顾名思义，就是对纽结的数学研究．图53(ⅰ)和(ⅱ)给出了两个简单的纽结例子：锁结和八字结．取一根绳子并打成上述的任何一种纽结，你将会发现：它们确实是打了结，而不是简单的“虚结”．区别在于：要打开一个纽结必须在某一步骤让绳子的一个自由端穿过绳圈，而一个虚结即使在绳子两端都固定的情况下也能被打开．另外，锁结与八字结似乎是很不一样的：要将其中一种结变换成另一种是不可能的，除非允许绳子的一个自由端穿来穿去．(如果两端都固定，那么任何操作都不能使锁结变成八字结．)

尽管到目前为止事情看似简单，其实我们已经履入薄冰．如果你[238]手里拿一根绳子，很可能经过半个小时的摆弄仍无法解开上面的纽结或将已有的结变成另一种结．然而这并不能证明绳子确实是打了

① 原文此处误作“球面”．——译者注．

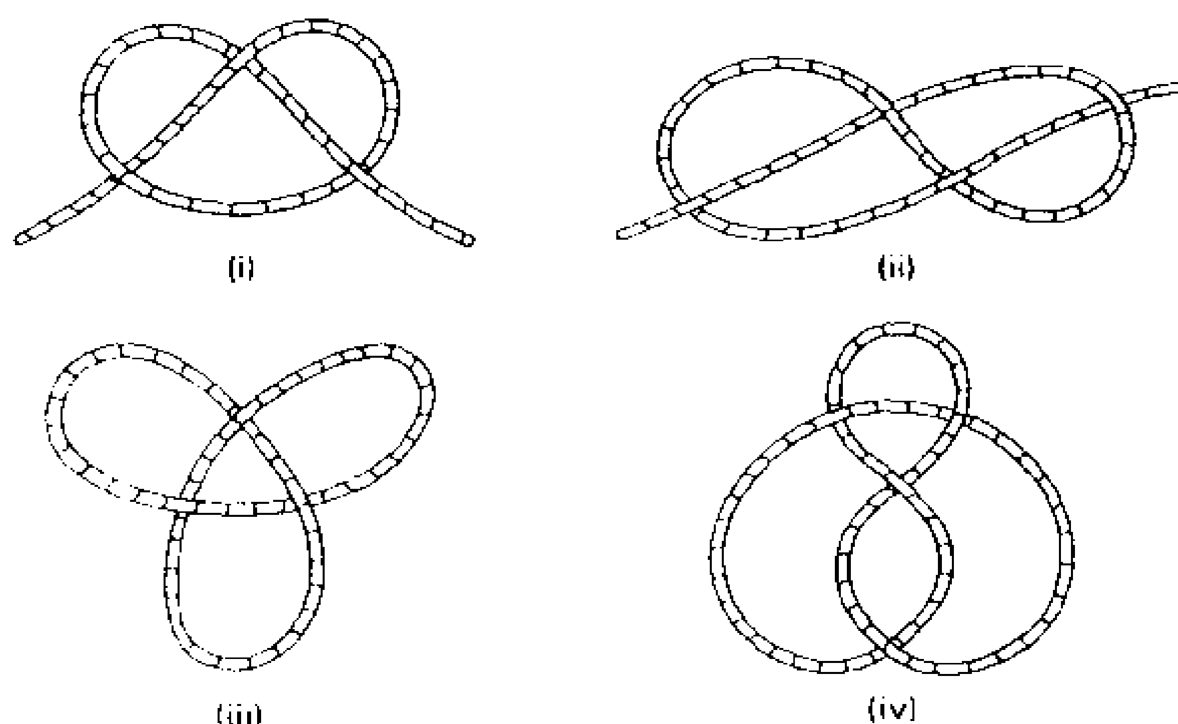


图 53 基本纽结:(i)锁结,(ii)八字结,(iii)三叶形纽结,(iv)八字形纽结.前两种纽结可以用绳线打成.但对数学研究来说,自由端必须连结起来形成一个闭圈.连结(i)的两端得到(iii);连结(ii)的两端得到(iv).图(iii)和(iv)就是纽结图的例子.

结或两种结确实不同.或许只是还没有找到正确的动作和步骤而已.(还记得图 48 中的难题吗?)对于复杂的构形来说事情又如何呢?当一长条绳线缠作一团(就像割草机的导线杂乱地缠在一起那样),要判别它究竟是一个虚结(在这种情形只要小心地抽拉线头就可将它展直)还是一个真正的纽结那是很困难的.一团乱麻似的线很可能根本就没有形成纽结,或者只不过是一个简单的纽结.魔术师们在舞台上就极尽玩弄"假结"的能事,那些糊人的假结其实根本就不成其为结.事情确实令人迷惑,而问题的解决需要依靠精细的方法——一种系统的、数学的方法.这正是纽结理论要研究的内容.

[239]

因为纽结理论的研究对象是"纽结性"(knottedness),于是首先要做的事情就是摆脱那些自由端.虽然实际打结的时候自由端非常重要,但它们的出现却妨碍数学研究的顺利进行,因为无论打结还是

解结都是(不含切割的)拓扑过程.换句话说,从拓扑学的观点看,带有自由端的绳子是不能打结的.最简单的解决办法是将自由端连结起来形成一个闭圈(可能已有纽结),在纽结理论中正是这样做的.如果将图53(i)和(ii)所示纽结的两个自由端连结起来,结果所得的纽结在数学上分别叫做三叶形纽结(trefoil)和八字形纽结(four - knot),如图53(iii)和(iv)所示.

因此,一个纽结就是一个闭圈(用绳、线或其他什么材料做成).(作为特例,这一定义包括了简单的"无纽结"的闭圈.无纽结的闭圈也叫平凡纽结,它有点类似于算术中的数零或集合论中的空集.)两个纽结被认为是等价的,如果有可能通过不包括切割的拓扑过程将其中一个变换成另一个.纽结理论最主要的目标就是要找出一组纽结不变量,依靠这组不变量就足以识别任意两个不等价的纽结(就像用可定向性和欧拉示性数来识别任意两个不等价的闭曲面那样).特别是应当能用这组不变量来确定一个给定的纽结是否确是纽结,或者仅仅是等价于一个平凡纽结(即一个简单的绳圈).

最早进行这方面研究的是在数学领地处处留芳的高斯.高斯的学生李斯廷(Listing)在其专著《拓扑学初探》(1847)中也用了很大的篇幅来讨论这一专题,并引发了一大批工作.(虽然直到1910年左右,德恩(Dehn)才试图证明确实存在非平凡纽结这样一种东西.在此之前则始终存在着理论上的可能性,即根本就没有需要研究的真正的纽结!)

1870 ~ 1900年间,由寇克曼(Kirkman)、泰特(Tait)、李特尔(Little)等人开展的早期工作,大多与所谓"素型纽结"(prime knot)的分类有关(可以想象,这种纽结不可能再被分解成两个更简单的纽结).分类的根据是这些纽结的所谓相交数(crossing number).为了得到一个纽结的相交数,首先将纽结在一水平面上摊平.对于用绳线做成的纽结来说,这是一个实在的操作过程,如果是一个概念的纽结,则可借助于"投影"来做这件事.然后设法使纽结本身尽可能少相交,并且在任何一点都不可能发生有三段或更多段绳线的交叉.(所谓相

交就是一个点，纽结的绳线在该处发生自身交叉.）绳线发生自身交叉的点的总数就是纽结的相交数.这是一个纽结不变量.例如图53(iii)和(iv)所示的纽结都已经按上述过程“摊平”，通过观察可以看出它们具有的相交数分别是3和4(并且都是素型纽结). [240]

作为一种纽结不变量，相交数可以被用来识别互不等价的纽结：如果两个纽结的相交数不同，它们就不可能等价.然而这概念却不如想象的那样有用.一方面，许多不同的纽结可以有相同的相交数，因此，单靠这种不变量往往不能发现不等价性.另一方面，相交数的计算十分困难，因为你必须用这样的方式来铺展纽结，使它没有不必要的闭圈和交叉，而这即使对相当简单的纽结而言，也不是显而易见的事情!

尽管如此，到19世纪末，已有大量相交数不超过10的素型纽结被分类制表.问题是这些表是否包括了每个相交数的所有纽结？表中是否有下述意义上的重复，即把表面不同但实质等价的纽结列成不同的项？1927年，亚历山大(J.W.Alexander)①和布里格斯(G.B.Briggs)研究了重复性问题，并设法证明了对于相交数不超过8的纽结来说不存在重复.他们的方法被应用于具有相交数9的纽结并且几乎获得了成功，只剩下3对纽结不能按他们提出的性质来加以区分.接着，瑞德马斯特(Reidemeister)对那三对特殊的纽结进行了研究.而对表中所有相交数为10的纽结的区别工作直到1974年才由彼尔科(Perko)完成.至于发现19世纪那张老表中所未能包括的新纽结的问题，则进展缓慢，直到1960年剑桥大学的J·H·康威(John Horton Conway)发明了一种新的更有效的纽结记号，才使他不仅能发现一些被前人遗漏的素型纽结，而且还将素型纽结表扩大到了相交数为11的情形.相交数不超过11的素型纽结似乎共有801个，其中

① 与传说“解开”著名的戈地结(Gordian knot)的那位亚历山大不是同一个人，后者是公元前333年的亚历山大大帝.他用宝剑斩断了那个著名的结，而这种方法在今天的纽结理论中是不允许的.——原注.

相交数为 3 和 4 的各 1 个，相交数为 5 的 2 个，相交数为 6 的 3 个，相交数为 7 的 7 个，相交数为 8 的 21 个，相交数为 9 的 49 个，相交 [241] 数为 10 的 165 个，相交数为 11 的 552 个①.（新表是否完全尚未最终证明，虽然人们相信是如此. 可能有重复的问题亦未解决，因此还不能保证上述这些数字都正确无误.）

关于纽结的数学理论，迄今最重要的进展是认识到如下的事实，即每个纽结都可与某个群联系起来，这群就叫（该纽结的）*纽结群*.（关于"群"的数学概念参阅第 5 章.）这是数学发展中激动人心的场合之一，这门科学某一部门的概念与结果在另一部门中获得了应用，[242] 眼下则是群论对纽结理论的应用.

构造纽结群的想法其实很简单，可以以三叶形纽结为例来解释（见图54）. 我们从选取不在纽结上的某一点 x 出发（点 x 具体选在何处并不重要，最终得到的群与此无关）. 为了定义一个群，现在要做三件事：(a)定义构成群的对象；(b)确定群中两个对象如何合成以产生第三个对象的法则；(c)验证每一条群的公理（见第104页）都成立.

就三叶形纽结的例子而言，构成群的对象是定向（即有箭头的）道路，这些道路从 x 点出发，不同程度地环绕纽结，最后又回到点 x（但它们不能穿过纽结的物质实体）. 并不是所有这样的道路都包含在群里，那些有多余闭圈（即在不影响道路环绕纽结的方式的前提下可被展开或割除的闭圈）的道路应去掉不计. 这样图 54 中的道路 g 就不在群内，道路 a 则在群内. 另外，若两条道路可以（在不"穿过"纽结实体的情况下）相互变换，如道路 a 和 b，则只取其中之一作为群的元素. 但注意 c 和 d 并不是这样的道路，因为它们以相反的方向环绕纽结，方向在这里是重要的因素. 有一条特殊的道路也属于群的

---

① 相交数为 3 和 4 的素型纽结分别如图 53 中(iii)和(iv). 对于直到相交数为 9 的素型纽结图参见 J. W. Alexander 和 G. B. Briggs 的论文"On types of knotted curves"（刊于 Annals of Mathematics, Volume 28(1927), pp. 562 – 86. 对于相交数为 10 的纽结，参见 K. A. Perko 的论文"On the classification of knots"（刊于 Proceedings of the American Mathematical Society, Volume 45(1974), pp. 262 – 6. ——原注.

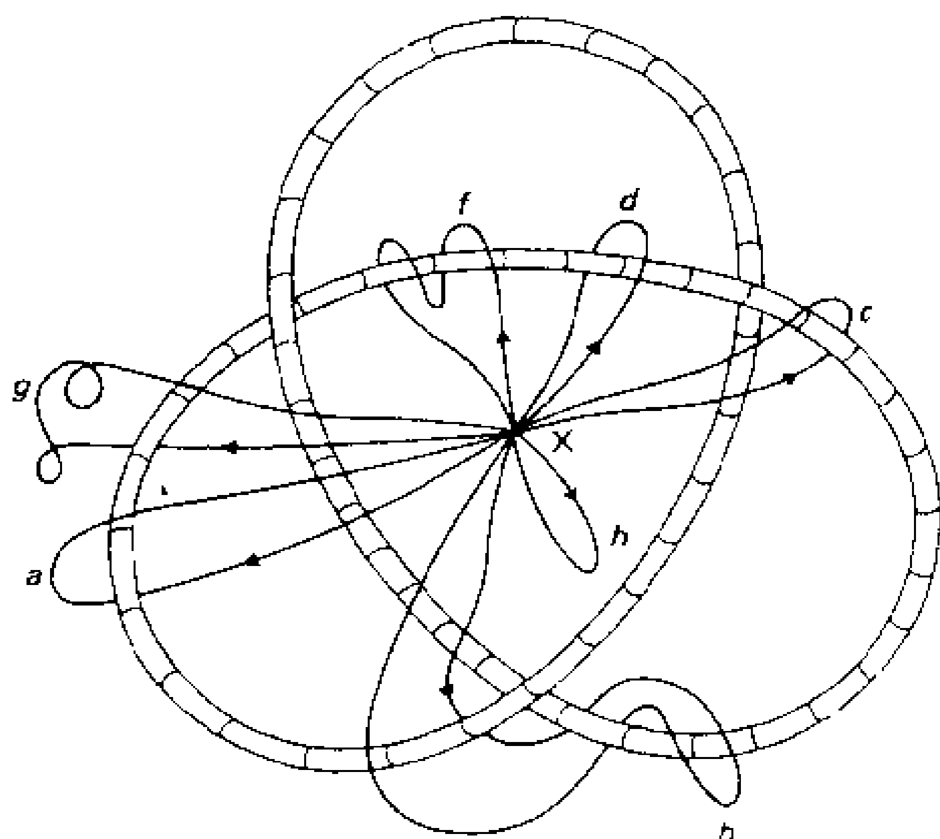


图54 三叶形纽结群的构造(详见正文).

元素,即长度为0的所谓平凡“道路”.任意一条完全不环绕纽结的道路,如道路 h,都可以变形为平凡道路,因此不是群的元素.这就是纽结群的定义.注意,这样定义的群包含有无穷多个元素,因为道路在回到 x 点以前可以任意无限多次地环绕三叶形纽结,而这些道路是各不相同的(即不能相互变换).

群的元素(它们与通常设想的群的元素颇为不同,这恰恰说明了群的概念范畴多么宽广)确定以后,下一步就是要定义两个这样的元素的合成或“乘积”.给定两条(定向)道路 p 和 q,先走 p 接着走 q 就合成另一条道路,我们总可以将这条合成道路变换成某条已属群内的道路,因为所有群外的道路都可以变换成群内的元素.这样的一条道路就是群内两个元素 p 和 q 的“乘积”$p * q$.例如,容易看出在图54中,$d * d = f$,在两条道路的合成中,相反的方向将相互抵消而不会留下任何道路.例如道路 c 和 d 就将相互抵消,若用 e 表示平凡道

[243] 路，这一事实就可以写成 $c * d = e$ 或 $d * c = e$.

剩下的事情就是要验证这样定义的系统满足群的公理.我们把它留给读者自己去做.单位元显然是平凡纽结 e，顺便说说，至少对三叶形纽结来说，纽结群是非交换的，这一点也让读者自己去证明.

纽结群是一种纽结不变量(这至少从定义来看是显然的，定义中已考虑到所有道路的变形)，两个等价的纽结将产生"相同"的群，就是说它们的纽结群就群的性质而言是相互的精确复制(exact copy).群中的具体道路看上去可以很不一样(两个有限群，如果它们的表(见第5章)完全相同，就说其中一个群是另一个的"精确复制".对像这里的无限群来说，为了使复制概念精确化，则需要利用同构(isomorphism)的概念).因此，纽结群像相交数一样可以用来区分互不等价的纽结.而纽结群则是更强有力的不变量.确实，纽结群更多地抓住了"纽结结构"的本质，很少有两个不同的纽结会具有相同的纽结群.(而大家熟悉的平结与错平结却是这样一对少见的纽结，这说明求知没有满足时!)因此，纽结群提供了区分纽结的有力工具，但问题在于：如何获得这种群的充分简单的代数描述呢？(请记住纽结群是一种无限的、抽象的数学结构.)利用上述关于纽结群的定义显然达不到这一目的，它比纽结本身还要复杂！幸运的是这个问题终于有了解答：1910年德恩发现了一种方法，可以从任一纽结的图形来得到相应纽结群的一种简明的代数表述.

这意味着纽结群确实提供了一种良好实用的对纽结进行分类的工具.它带来了强有力的群论方法，有时还引出其他有用的不变量的发展(我们将在稍后介绍其中的一种——亚历山大多项式).这里同时介绍一下纽结群的一个简单而直接的应用.试考虑以下问题：给定一个纽结，能不能用这样的方法将它解开(所谓解结就是使之与圆等价)，即先附加上一个纽结，然后用它去抵消原有的纽结？回答是否定的.附加纽结将会产生一个比原先更大、更复杂的纽结群；与此同[244]时，平凡纽结(即圆)的纽结群则与整数关于加法运算而形成的群相同.因为纽结群不同，纽结本身也就不可能等价.

另一条研究纽结的途径是1935年由谢菲尔德(Seifert)开辟的.他设计了一种对任意给定纽结构造一个以此纽结为唯一边界的可定向(即双侧)曲面的方法.当前,曲面拓扑学(见下节)的一个标准结果是:任一单边、可定向曲面都拓扑等价于具有一定个数"环柄"(handle)的圆盘.在圆盘上开两个洞,然后将一根圆柱管从一头拉到另一头,这就是加有环柄的圆盘(见图59).一个圆盘的环柄数叫做原来曲面的亏格(genus),亏格与欧拉示性数一样也是曲面的一种拓扑不变量.这样就有可能利用谢菲尔德方法使任意纽结与一个自然数联系起来,这个数就是该纽结的谢菲尔德曲面的亏格,也叫该纽结的亏格.(有一点小小的麻烦是:从同一个纽结出发,用谢菲尔德的方法可以得到不同的曲面——即有不同的亏格.你要取的是用这样的方法从纽结得到的最小亏格.)

像纽结群一样,亏格也能够被有效地用来判别一个纽结是否真正有结.平凡纽结的亏格显然是零,因为由圆围成的圆盘没有环柄.另外,平凡纽结是唯一亏格为零的纽结,当然这一点不如上述事实那样明显.因此为了证明一个纽结确实是结,只要证明其亏格不等于零就行.可是,给定了一个纽结图后,怎样来具体计算它的亏格呢(这正是你常常要面临的问题)? 1962年,哈肯(W.Haken,因四色定理而著名,见第7章)提出了一种可能的方法.1978年,赫米翁(G.Hemion)利用哈肯的结果构造出一种算法,借此可判别两个给定的纽结图是否表示相互等价的纽结.遗憾的是这个算法效率太低,缺乏实用价值(关于算法有效性的讨论参阅第11章).不过它确实说明了纽结分类问题原则上可以用机械化的方法来解决.

虽然谢菲尔德的方法采取了几何而不是代数的途径.1978年,福斯泰尔(C.Feustel)和惠顿(W.Whitten)却指出了如何从纽结群来获得亏格的方法,从而又一次强调了纽结群的基本性.

读者现在大概已经被为了区分纽结而引进的一大套复杂的数学方法弄得晕头转向了.你可能会问:"难道就没有更简单的方法吗?" [245] 如果你勇于经受偶尔的失败(个别情况下不能区分不等价纽结),那

么还真有一种更简单的方法．这就是所谓的纽结多项式方法．最简单的纽结多项式是亚历山大多项式，1928 年由亚历山大发现．虽然一个纽结的亚历山大多项式可以从纽结群推导出来，但它也可以直接通过纽结图而得到．平凡纽结的亚历山大多项式是数 1，三叶形纽结（见图 53(iii)）的亚历山大多项式是

$$x^2 - x + 1.$$

而八字形纽结的亚历山大多项式则为

$$x^2 - 3x + 1.$$

显然这两个多项式是不同的，所以三叶形纽结与八字形纽结互不等价（它们也都不是平凡纽结）．纽结的亚历山大多项式是一种很有用的不变量：它足以用来区分相交数不超过 8 的所有素型纽结，并且除了 6 对例外，还可以识别所有相交数为 9 的素型纽结．

然而，在亚历山大多项式区分不了的纽结中，有一对就是平结和错平结．二者都有相交数 6，并且都不是素型纽结，每个都由两个三叶形纽结合成．区别在于：构成它们的三叶形纽结是左手结还是右手结（见图 55）．这两个纽结的亚历山大多项式都是

$$(x^2 - x + 1)^2.$$

（恰好是三叶形纽结亚历山大多项式的平方，这并非偶然，任何一个合成纽结的亚历山大多项式都是构成它的纽结的亚历山大多项式的乘积．）不过请不要为此而怪罪亚历山大多项式．正如前面所说，即使是更强有力的纽结群也不能区分这两种纽结．问题的焦点在于左手三叶形纽结和右手三叶形纽结的区分．纽结群做不到这一点，至少单靠它自己不行．用适当的方式"加强"纽结群，则有可能得到一种胜任此项任务的纽结不变量，但这又要涉及更加抽象的方法了．

普通的童子军轻而易举的事情，却令数学家们搜尽抽象方法的武库．直到 1984 年 8 月，事情才有明显的转机．这一年有四个相互独立的数学家小组（一个在英国，其他三个在美国）发现了一类新的多项式，用这类多项式就可以识别左手三叶形与右手三叶形的纽结（也就是说可以识别平结与错平结）．按这类多项式发现的情况，如果用

[246]

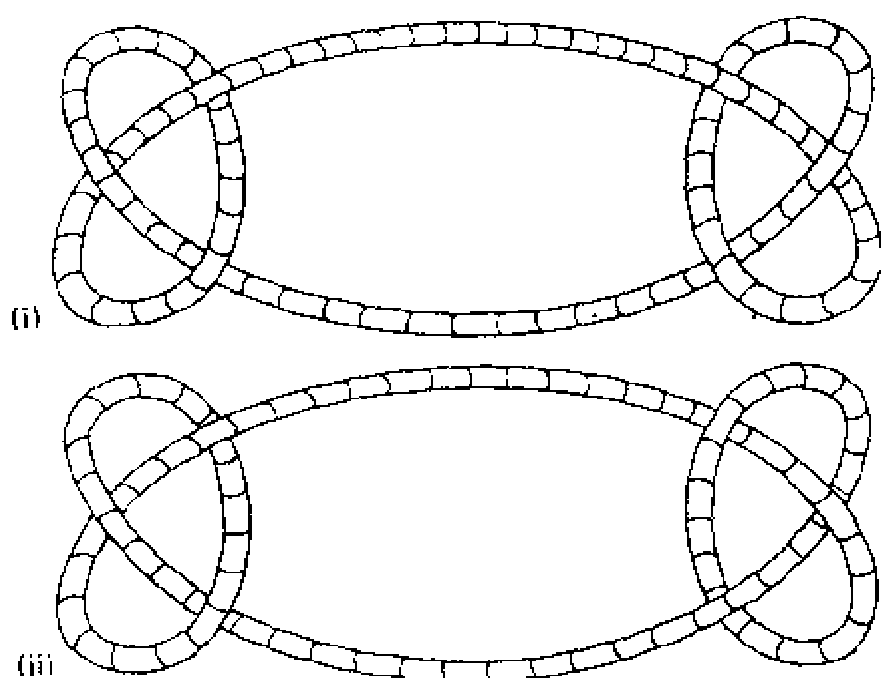


图 55 (i)平结,(ii)错平结.二者的区别在于其中三叶形纽结的走向.纽结群不能区分这两种定向.

有贡献的数学家的名字来命名的话,可称其为康威-琼斯-弗雷德-叶特尔-霍斯特-李柯里希-米勒特-奥克尼努(Conway-Jones-Freyd-Yetter-Hoste-Lickorish-Millett-Ocneanu)多项式,念起来可真拗口.幸好多项式本身却简单得多.与只有一个变元的亚历山大多项式不同,这种新多项式有两个变元.(它们还包括有这些变元的负次项,因此读者一开始可能会犹豫是否该称它们为"多项式".)右手三叶形纽结的新多项式是

$$-x^{-4}+x^{-2}y+x^{-2}y^{-1},$$

左手三叶形纽结的新多项式是

$$-x^{4}+x^{2}y+x^{2}y^{-1},$$

[247]

而八字形纽结的新多项式则是

$$x^{-2}+x^{2}-y-y^{-1}+1.$$

那么怎样计算这些多项式呢?它们可以从纽结图得到.具体计算过程已超出本书范围,但应该指出这是一种可以有效地在计算机上进

行的彻底机械化的过程.

这种多项式能够区分所有的不等价纽结吗？不能,它们尚不能胜任.数学家们还必须继续探索.纽结理论中还有大量未被解开的结.但对我们来说,现在应该转向另外的话题了.

在曲面上挖洞①

自从50年代以来,拓扑学发展的中心课题无疑是流形(manifolds)的研究.粗略地说(恰当的定义将在后面给出),流形乃是曲面概念在任意维情形的推广.最简单的流形是一维和二维的流形.一维流形就是曲线(实数轴$\mathbb{R}$是其特例),而二维流形即曲面(二维平面$\mathbb{R}^2$是其特例).二维流形的好处是可以画出它们的图像(甚至还可以作它们的物理模型).对本书读者来说,遗憾的是实际上自本世纪初以来,几乎所有的研究都集中在三维或更高维的流形,而这类流形是无法用图形来解释的.(一般书本中与高维流形有关的图像都是画给专家看的,需要做详细的解释.)当然,如果维数不高,例如在三维或四维情形,还有可能用投影或取截面的方法来解释简单的流形.但想弄懂它们,仍需要加些附带的说明.例如,如果没有解说词,你能认出图56所要表示的是什么样的四维形体吗？这是一个超立方体(即立方体的四维类似物)的两张投影图.正像一个三维形体可以通过在二维[248]平面上的投影来解释一样,一个四维形体,原则上也可以通过它在三维空间中的投影来描绘.再将三维图形投影到平面上,就能得到原来四维形体的一个图解,这就是在图56中所看到的东西.当然,解释这些投影需要费一番脑筋.将三维形体投影到平面上做起来比较容易.荷兰画家埃歇尔(Mauritz Escher)一些引起轰动的作品就是利用这样

① 原文"*Scratching the surface*"亦作英语成语"浅尝辄止",此处语带双关,因本节是以曲面(二维流形)为例(在曲面上挖洞)来为理解高维流形理论作引导.——译者注.

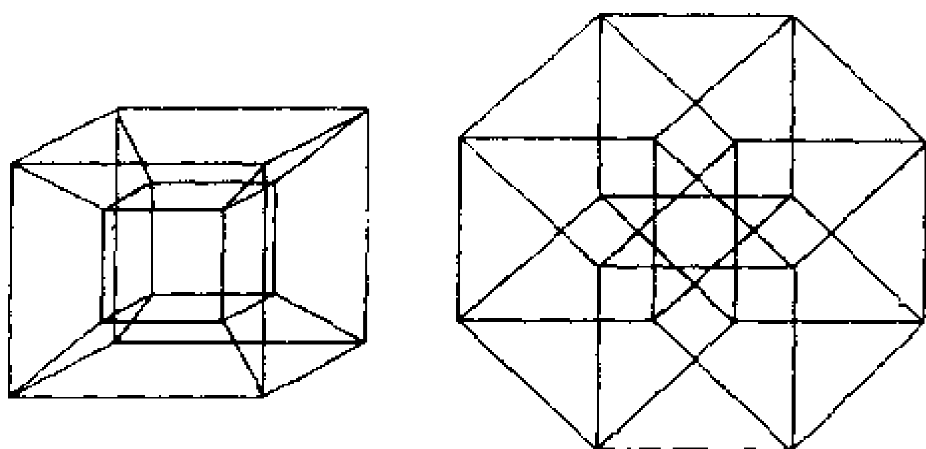


图56 超立方体,一种具有8个大小相同的正方体作"面"的四维图形.两张图都是表示超立方体在三维空间中投影的平面投影.

的技巧制作出来并收到艺术效果的,图57那样实际上"不可能"的绘画就是一例.四维形体的投影解释起来就比较困难——超立方体究竟是什么样子呢?正像一个通常的立方体有6面,每个面都是大小相同的正方形一样,一个超立方体有8"面",每个面都是大小相同的正方体.如果你仔细观察一下图56左边的图(同时试着想象它所描绘的三维投影),你将看出它表示了8个正方体:最外面的那个大的、最里面的那个小的,以及已变形为平截头方锥的那6个.大小与形状的畸变只是从四维向三维空间投影的特征,在实际的(四维)空间中,这8个"面"大小都是相同的,而超立方体则"介于"它们之间.

除了投影,另一种使高维形体形象化的方法就是取"截面".例如,什么样的四维形体会给出图58所示的一系列三维截面呢?我们[249]知道一个球的相继截面形成一系列的圆,这些圆从最小的开始,逐渐变到最大,然后又重新变小.(这类似于将苹果切片的情形.)同样的[250]一个四维超球的相继截面将形成一系列的球,如图58所示.(问题:一个超立方体的一系列截面是什么呢?)

这样的图解方法至少可以使人对四维物体是何样有个粗略的概念.它们对说明五维或更高维的形体却毫无用处.不过,它们指出了一条通过逐步类比而从低维向高维进展的卓有成效的途径.例如,

图 57　M·C·埃歇尔的艺术作品：升与降(1960)．运用从二维投影解释三维形体的技巧而得到的实际上不可能的图形．

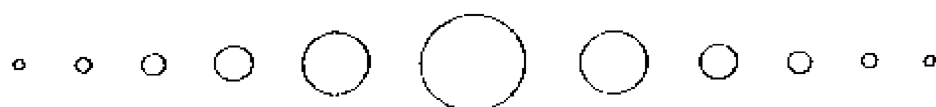


图 58　一个四维超球的截面图．如果你能够将一个四维苹果切片，得到的就是这样的东西：一系列由小变大又重新变小最后消失的球"片"．

个二维正方形的界"面"是4条等长的一维线段,一个三维立方体的界面是6个大小相同的二维正方形,一个四维超立方体的界"面"是8个大小相同的三维正方体等等,依次类推.

但请注意！事情并不像初看那么简单.正如希腊几何学家已经知道的那样,在三维空间中总共只有五种正多面体:正四面体,正立方体,正八面体,正十二面体和正二十面体.在高于三维的空间,多面体的类似物叫作多胞形(polytope,在第11章中还将提到这一概念并讨论其对三维现实世界的应用).一个四维多胞形的"面"是一些三维多面体.对一个正多胞形来说,这些"面"本身必须是正多面体,同时在每个顶点处有相同的排列.现已弄清,四维多胞形也只有六种:具有五个正四面体"面"的单形(simplex),具有八个正方体"面"的超立方体,由16个正四面体围成的16胞形(16-cell),有24个正八面体"面"的24胞形,有120个正十二面体"面"的120胞形以及有600个正四面体"面"的600胞形.因此,当你从三维进到四维的时候,事情变得复杂了.但接着却发生了奇怪的事情.对于任意维数大于4的空间,其中的正多胞形都只有3种,它们分别类似于正四面体,正方体和正八面体.那么还会发生什么情况呢?为什么超过四维以后事情反而突然变得简单(而没有变化)了呢?虽然还没有人能回答这个问题,这种现象却并非正多胞形所特有,在其他许多方面,五维或更高维空间也要比三维或四维情形容易处理.

251

尽管在高维情形有某些新的因素在起作用,考虑一下二维流形(曲面)的行为仍然会对了解流形理论所涉及的问题与方法的本质提供合理的思想,因而曲面的拓扑理论值得进一步探讨.

前面提到的二维流形的分类是19世纪拓扑学最重大的成果之一.为了对所有的闭曲面加以区别,只需要两个不变量:可定向性与欧拉示性数.怎样来实现这种分类呢?现在的证明通常分成两个阶段.首先是证明每个闭曲面都可以拓扑变形为两类标准的形式之一.然后再证明依靠可定向性与欧拉示性数这两个不变量已足以区分所有的标准曲面.

具有亏格 n 的标准可定向曲面是由一个球面与连贴在球面上的 n 个环柄组成. 为了将一个环柄连贴到球面上, 只要在球面上开两个洞, 然后再用一段圆柱管将它们连结起来(参见图 59). 带有任意个环柄的球面都是可定向曲面. 一个有 n 个球柄的球面其欧拉示性数为 $2-2n$. 证明这一点并不难——方法是从球面上的一个网络开始(对此网络有 $V-E+F=2$), 然后增加其环柄. 认真地做下去你就会发现: 每增加一个环柄, 欧拉示性数就减少 2①.

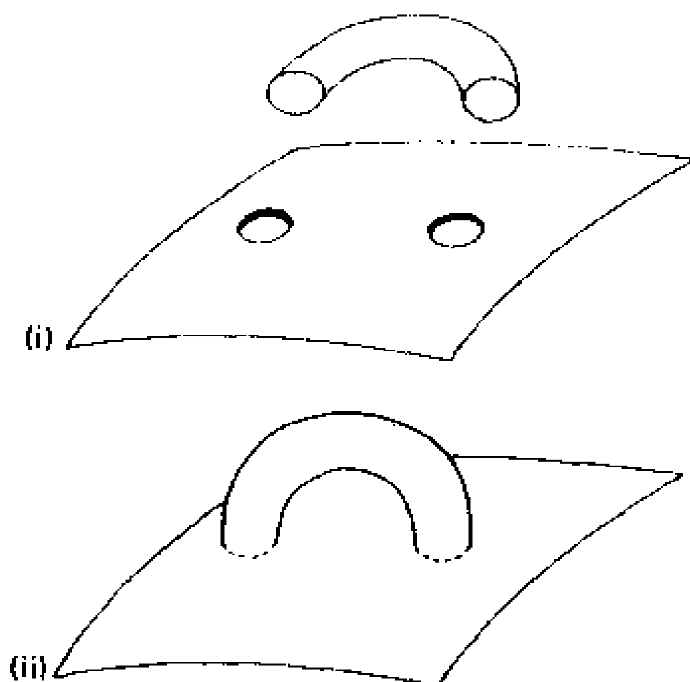


图 59 环柄. 在曲面上加一个环柄, 首先是如图(i)所示在曲面上开两个洞, 然后如图(ii)所示在两边缝上一圆柱管将它们连结起来. 使用所谓的"换面术"可以证明: 每个可定向闭曲面都拓扑等价于具有一定个数环柄的球面, 这就给出了可定向闭曲面的"标准型".

通过由分割、拉开和重新装配组成的过程, 不难把任意的可定向曲面变形为具有某一亏格的标准可定向曲面, 这样的过程由于明显的理由而被称为换面术(surgery). 例如, 环面表示亏格为 1 的可定向

① 这一过程实际上也就是整个分类证明, 其详细介绍可参看 Ian Stewart: Concepts of Modern Mathematics (Penguin, 1981) 一书第 12 章. ——原注.

曲面，双环面表示亏格为2的可定向曲面，等等. 因为对可定向曲面来说亏格与欧拉示性数是相关的(通过上述表达式 $2-2n$)，这就说明欧拉示性数是怎样被用于所有可定向曲面的分类的. [252]

亏格为 n 的标准非定向曲面可通过在球面上加 n 个叉帽(cross-cap)来得到. 所谓在一个曲面上加一个叉帽，就是在这曲面上开一个洞，然后缝上一个莫比乌斯带并使边缘与边缘连结在一起. 在三维空间这很容易做到，只要你允许莫比乌斯带自身相交(参见图60)，因为在莫比乌斯带上顺时针方向可以变换为逆时针方向，所以有一个叉帽的曲面是非定向曲面. 有 n 个叉帽的球面的欧拉示性数为 $2-n$.

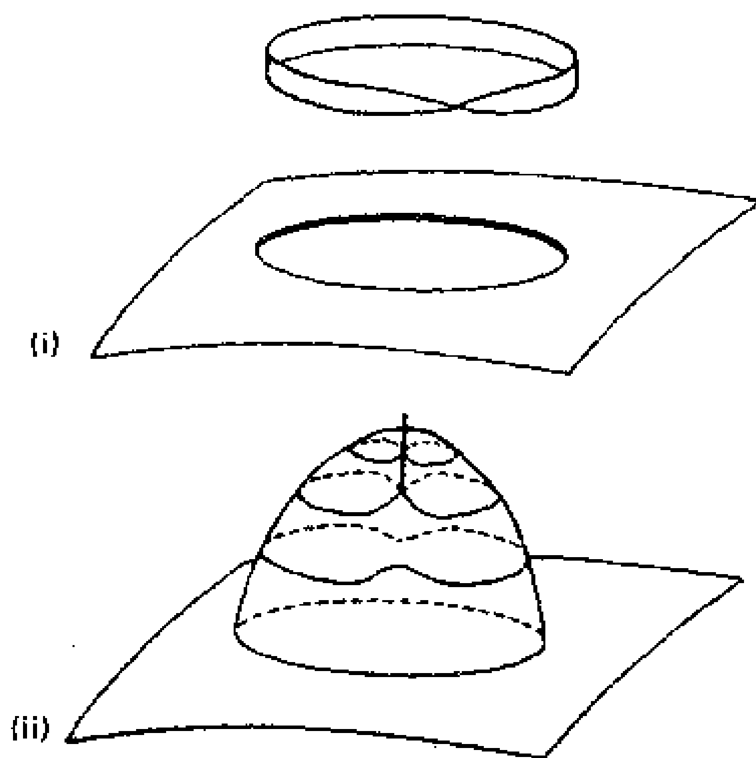


图60 叉帽. 在曲面上加一个叉帽，首先是如(i)所示在曲面上开一个洞，然后边对边缝上一个莫比乌斯带，这在三维空间很容易做到，只要允许莫比乌斯带自身相交(如(ii)所示)，结果就得到一个叉帽. 使用拓扑换面术可以证明：任意非定向曲面都拓扑等价于具有一定个数叉帽的球面，这就得出了非定向闭曲面的“标准型”.

这同样也可以通过下述方法来证明，即在球面上取一网络并说明每增加一个叉帽时数量 $V-E+F$ 减少 1(详细介绍可见上述 Stewart 的书).

借助于换面术，有可能(且并不特别困难)使任意的非定向曲面变形为具有某一亏格的标准的非定向曲面. 例如，几何学家的投影平[253]面(尽管名曰平面，实际上却是一个闭曲面)可以变换成一个亏格为 1 的标准非定向曲面. 克莱茵瓶则可变换成一个亏格为 2 的标准非定向曲面. 因为标准非定向曲面的欧拉示性数与亏格相关(通过表达式 $2-n$)，上述标准化过程同样也是利用欧拉示性数而建立了所有非定向曲面的分类. 这里涉及了有边缘的曲面，因为要允许在标准曲面上开洞.

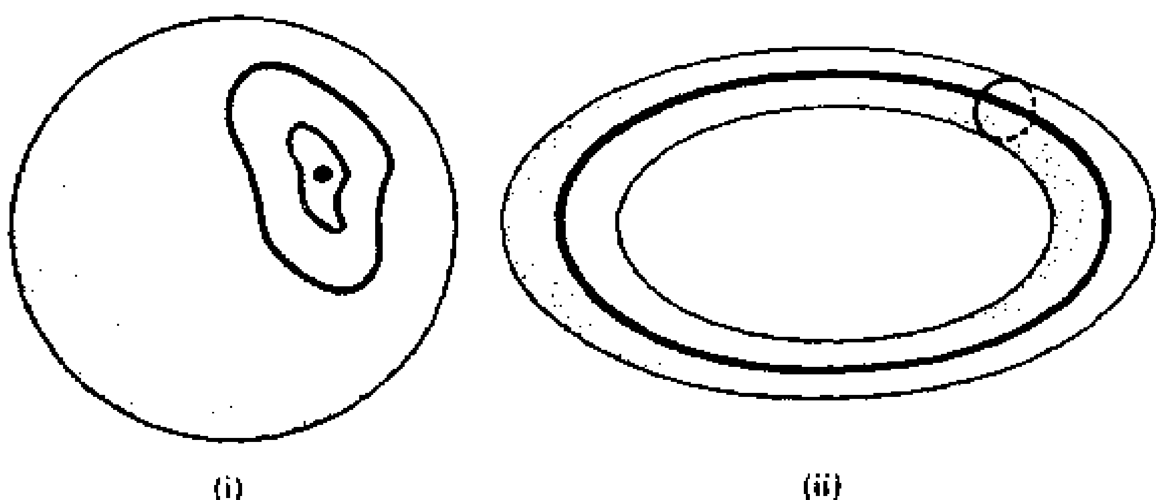
上述曲面拓扑的初步知识使我们能够去进一步了解近年来高维空间研究的进展. 我们从考虑一种最简单的流形即 n 维球面($n=2$, [254] $3,4,\cdots$)开始. 最著名的拓扑学问题就是与这样一些流形有关的.

庞加莱猜想

在所有 2 维闭曲面中，最简单的是球面，上述的分类过程就是从球面开始的. 球面在 n 维空间中的类似物称之为 n 维球面. (这样通常所说的球面就是指 2 维球面.) 正像 2 维球面是一个 3 维球体的表面一样，n 维球面也是一个 $n+1$ 维"球体"的表面. 当本世纪的法国数学家庞加莱开始研究高维流形(实质上就是我们今天所理解的流形拓扑学的先声)时，他很自然地对 n 维球面特别关注. 这些球面应该是很有特殊性的，正如 2 维球面在所有的 2 维流形中具有特殊性一样. 1904 年，庞加莱试图证明一个他认为对 n 维球面来说是理所当然的推断，结果未能如愿. 于是他便将这推断作为猜想提出，庞加莱的这个猜想已成为流形拓扑领域中最著名的问题. 像所有好的猜想一样，它是基本的，同时又是容易陈述的.

假如在 2 维球面上画一个闭圈，在不离开球面的情况下，那么使

这个闭圈缩成一点是可能做到的(参看图 61(i)).不仅如此,球面还是具有这种性质的唯一的闭曲面.如果在一环面上按图 61(ii)所示两种方式中的任何一种去画闭圈,你就不可能使它缩成一个点.类似地(只有现在,你才须依靠抽象数学而无法借助于图形),设在一 n 维球面($n>2$)上"画"一闭圈,则这闭圈也可以在不离开 n 维球面的情况下缩成一点.然而一个严重的问题是:n 维球面是否也像 2 维情形一样是唯一具有此种性质的 n 维闭曲面呢? 庞加莱猜想断言:答案是肯定的(严格地说,庞加莱只对 3 维流形感兴趣).



(i) (ii)

图 61 庞加莱猜想.在球面上画一个闭圈,在不离开球面的情况下将这个闭圈缩成一点是可能的,如(i)所示.在环面上却不一定能这样做.如图(ii)中所示的环面上的两个闭圈,都不可能缩成一点而又不离开环面.对闭曲面来说,这种能将任一闭圈缩成一点的性质完全是球面的特性:其他所有的闭曲面都不具有这种性质.庞加莱猜想是说:类似的结论对所有的高维情形也成立.例如,唯一具有缩圈性质的三维闭流形就是三维超球面.

尽管数学家们绞尽了脑汁,1960 年以前证明(或否证)庞加莱猜想的所有尝试都归于失败.直到 1960 年,美国数学家斯梅尔(Stephen Smale)首次证明了:庞加莱猜想对于 5 维和高于 5 维的情形成立.斯梅尔由于此项工作而当之无愧地被授予菲尔兹奖章.然而

[255] 斯梅尔的方法对于3维和4维的情形却显得无能为力，这又一次为我们前面已经提到过的现象提供了例证：即流形的行为从5维开始变得有所不同. 事实上，又花了将近20年的光阴，四维问题才被另一位美国数学家解决. 1981年，弗里德曼(M. Freedman)在斯梅尔思想和坎松(Andrew Casson)工作的基础上证明了四维庞加莱猜想. 弗里

发展任意高维流形$\mathbb{R}^n$($n=3,4,5$,等等)上的微分学.

因为任意n维流形都与$\mathbb{R}^n$局部相像,当然就可在流形的局部区域(local basis)上使用通常的微分方法.但整体怎样呢?至少对球面来说有可能建立覆盖整个曲面的微分运算.原因是这里从一个局部区域(与$\mathbb{R}^2$相像)到另一个局部区域的过渡是光滑的和没有困难的.换一种方式来说,设用经线和纬线来覆盖整个球面,则在一局部基础上就可得到一个坐标系,它类似于通常的笛卡儿几何坐标系(后者正是微分学赖以发展的背景).如果你采用这些坐标来建立球面上的局部微分学,那么由于在整个曲面上使用的是同样的坐标线,在一个局部区域上所发生的事情与另一局部区域就不会相互冲突:所有的过渡都将是光滑的. [257]

于是发生了一个自然而基本的问题:究竟有多少流形可像在球面上那样建立适合整体的微分运算?可以建立整体微分理论的流形叫作光滑流形(有时也叫微分流形).而可覆盖整个流形并作为微分过程的基础的坐标系(类似于球面上的经、纬线)被称作"微分结构".(事情实际上更为复杂,但这多少提供了一幅形象图.)上述基本问题就变成:什么样的流形是光滑流形(或者说什么样的流形可以被赋予微分结构)?伴随这基本问题的另一个同样有趣的问题是:给定一光滑流形,是否只有一种还是可用多种方式赋予其微分结构(如有多种方式,则问究竟有几种?),因为物理学家们要花费大量的时间去进行各种流形上的微分演算,回答这些问题就不仅仅是拓扑学家感兴趣的事情了.

对二维或三维流形,上述问题的答案在50年代中期就已经知道!每一个二维或三维流形都是光滑的,并且任何这样的流形都不可能有两种本质不同的微分结构.当时人们以为上述结果向任意维数的推广似乎只是时间问题.然而,1956年美国的米尔诺(John Milnor)出人意料地发现:7维球面可以被赋予28种不同的微分结构.不久人们又发现其他一些高维球面也不只有一种微分结构.显然,进一步的探究还有大量的工作要做,而跃跃欲试的有能力的数学家也大

有人在.从 1956 到 1970 年这一时期被称作“流形拓扑学的黄金时代”(实际上应该说是 5 维和 5 维以上流形研究的黄金时代,因为 4 维问题在这里再次被证明难于用已有的方法来处理).正是在这个时期,拓扑学家们通过使用所谓“同伦”的概念获得了所有高于 4 维的流形的系统分类,特别是得以对光滑与非光滑的流形作出区分.

那么 4 维的情形如何呢?是否也像低维情形那样,所有的流形[258]都是可微的并且只容许一种微分结构?或则可能像高维情形那样需要加以分类?最终答案是在 1981 年得到的.弗里德曼在解决 4 维庞加莱猜想(如前所述)的同时证实了存在有非光滑的 4 维流形(这种流形由于技术上的原因而称为 E_8,弗里德曼的描述如同高维拓扑的所有其他事情一样是代数性质的).实际上,4 维庞加莱猜想和非光滑 4 维流形的发现都是弗里德曼获得的一个非常一般(并且完全出乎意料)的结果的推论,此结果表明:为了对任意 4 维流形分类,只需要两项“初等”的信息就够了.(但这些信息实际上并不那么“初等”,因此不可能在这里解释.)

可是这并不是全部故事的结局.还有另外的、同样令人震惊的戏剧性结果等待着拓扑学家,并且事情很快就发生了,这次惊人的结果恰恰切中了与我们所生活的物理宇宙有关的核心问题!

关于流形的出人意料的结果常常可以这样来解释搪塞,即人们处理的是一些漂亮抽象的概念,而这些概念至多只能得到部分理解.即使是二维流形也可以是漂亮的虚构之物.但对 $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{R}^3$ 等“具体”的流形我们却不能作这样的解释.因为 $\mathbb{R}^3$ 毕竟是我们生活的物理空间(果真如此?),而 $\mathbb{R}^4$ 则是时-空连续统.这些“具体”流形也确实表现出典型的行为.首先,它们都是光滑流形.其次,对每个 n, $\mathbb{R}^n$ 都只能以一种方式被赋予微分结构——只有一个例外的情形 $n=4$.

由于某种特殊的原因,数学家们至此未能证明 $\mathbb{R}^4$ 上微分结构的唯一性.对其他任何 $\mathbb{R}^n$ 都能证明,唯独 $\mathbb{R}^4$ 不能.尤为恼人的是 $\mathbb{R}^4$ 正是物理学家们最感兴趣的情形.或许发现证明仍然只是时间问题.在 $\mathbb{R}^4$ 上可以有非标准的微分方法那简直是不可思议的……?

然而不可思议的事情竟成为现实.1982年夏天传来了爆炸新闻.M·阿蒂亚(Michael Atiyah)24岁的学生,牛津大学的S·唐纳尔逊(Simon Donaldson)把弗里德曼的(本质上是代数的)工作与大量的分析与微分几何工具相结合而证明了一个结果,由此结果可推出$\mathbb{R}^4$上与通常不同的微分结构的存在性.换句话说,世界物理学家和数学家们所惯用的微分结构并不是唯一的!(事实上,C·托勃斯(Clifford [259] Taubes)随后的工作说明$\mathbb{R}^4$上通常的微分结构只不过是可以赋予该流形的无数微分结构中的一种!)这就产生了两个令人费解的问题.究竟是什么样的特殊性使上述现象仅仅发生在4维情形?因为在$\mathbb{R}^4$上可以有多种微分方法,我们能否判别哪种方法对物理世界来说是正确的?由于所有n维流形除n等于4以外都已有章可循,$n=4$的情形就变得越来越奇特了.

关于$\mathbb{R}^4$,物理学家的工作中使用的是正确的数学吗?也许是.

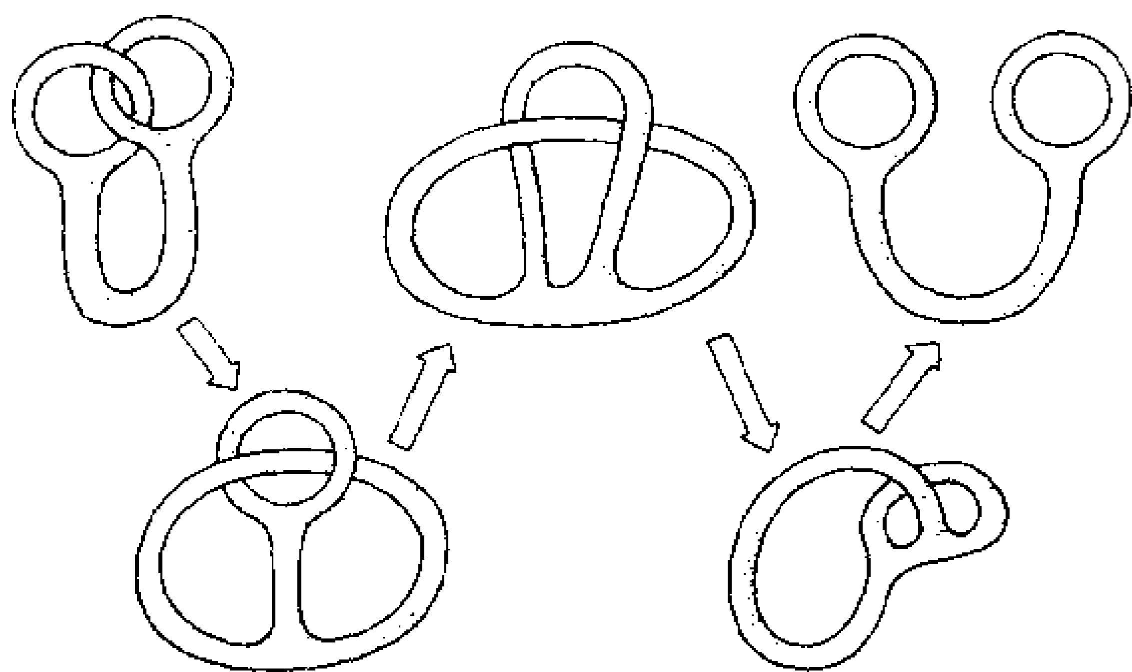


图62 环的难题(图48)之解决.图中所示过程说明了将原先连在一起的环构形变成分开的环的步骤.

迄今发现的$\mathbb{R}^4$上无穷多种“外来”的微分结构全都表现出某些特异古怪的行为，从而不适用于我们的物理宇宙.但它们的存在说明了四维空间确实具有某些非常特殊的性质.它们的发现标志着我们确实面[260]临一个拓扑学的黄金时代.

阅读文献

关于拓扑学的比本文更全面但程度基本相当的介绍参阅I.Stewart的Concepts of Modern Mathematics (Penguin,1981),10－14章.

供外行阅读的介绍纽结理论的文章可参阅Lee Neuwirth的The theory of knots,刊于Scientific American, Volume 240(June 1979),pp.84－96.更全面(水平更高)的介绍则可参阅R.Crowell和R.Fox的Introduction to knot theory (Ginn and Company, USA, 1963).关于这一领域的一个出色的综述参阅M. Thistlethwaite的文章Knot tabulations and related topics,刊于I.M.James和E.H. Kronheimer合编的文集Aspects of Topology(London Mathematical Society Lecture Note Series, Volume 93(Cambridge University Press, 1985)).

关于流形理论,比上面提到的Stewart的书程度更高的著述对外行来说很难读懂.为了能得到一些概念,读者似乎可浏览一下D.Freed和K.Uhlenbeck合著的Instantons and Four－Manifolds(Springer－Verlag, 1984),该书重点介绍了Donaldson和其他人最近关于非标准$\mathbb{R}^4$流形的工作.(但仅当你已经是一位代数拓扑专家时,才能考虑买下这本书来详读.如果你不是这方面的专家,那么本书[261]就不是为你写的.)

(李文林译)

第11章 算法有效性

再谈算法

算法的概念在本书第6章中已扮演过重要的角色(从现在起假定读者已读过该章).那里所讨论的希尔伯特第十问题是问:某个特定的问题是否能有算法求解?这里强调的是"有".纯粹的存在性是当时的要求——没有人关心所讨论的算法是否实际可行.就希尔伯特问题而言,这当然已完全符合要求.但当我们面临周围现实世界所提出的问题时,仅仅知道一个算法存在决不能算是事情的结束.实际上这仅仅是开始,因为从一个算法的存在并不能得到什么实际的好处,这个算法虽然可以在理论上解决你手头的问题,但却有可能让一台高速计算机算上几千年.对于商人和应用科学家们所关心的各种问题来说,重要的是一个有效算法的存在性.在商业活动中,问题的解答有时需要在几小时之内得到,而像飞机导航这样的系统则往往需要在几分之一秒钟内就能算出结果.对诸如此类的应用而言,证明某个特殊的问题能不能用"有效算法"来解决显然十分有用.为此,第一步就是要提出一种评估算法有效性的适当方法.

显然,在计算机上解决一个给定问题的速度依赖于一系列因素. [262] 计算机的规模与运行速度,编写程序的语言的效率、程序员的熟练程度等等都与此相关.不过这些都是很特殊的因素,不属于一般研究的范围.我们需要的是某种非常一般的方法,利用它来将算法分成两大类:有效算法与非有效算法.这样的分类应该充分固定,以致无论怎样

更换诸如计算机速度或程序语言这样一些因素，都不可能使非有效的算法变成有效算法，或者反之使一个有效算法变成非有效的算法.

这种有效性分类方法是柯勃汉(A.Cobham)和爱德华兹(J.Edwards)在60年代中期首先引进的，现已成为绝大多数关于算法有效性工作的基础. 虽然他们是用"时间"来作为有效性的基本度量，但为了避免对计算机速度的依赖，实际定义是借助执行计算时所需的步数来给出的. 当然，即使这样的定义也不是绝对的——它可能依赖于什么是基本步骤和怎样来表示数据. 但后来人们终于明白：就有效性的基本概念而言，这些顾虑是多余的. 因此借助图林机(参见第6章)来表述各种定义就成为标准的做法. 这种做法非常简单，足以形成漂亮的数学理论，而且不管你用什么样的计算机来重演计算过程，所得的结论都不会改变.

确定以图林机作为基本的计算工具，目的是要用完成计算所需的步数(图林机上的步)来衡量一种算法的有效性. 至于怎样将这个算法编成图林机程序，以及怎样在机带上表示数据，这样一些问题已被证明是无关宏旨(也就是说这些因素并不影响有效算法与非有效算法的分界). 有关系的是必须处理的数据的多少. 数据越多，所需的处理步骤也就越多. 例如用手算来进行一对整数的乘法，如果把这对整数的字长加倍，那么所需的处理步骤就至少增为4倍，因为需要进行的基本数字乘法运算次数是原先的4倍，另外还要加上记录历次进位所需的"超支". 记住这一点，我们就可以给出如下的一些定义.

一个算法(在这里的定义中即一个图林机程序)被称为是多项式时间算法，如果存在确定的整数 A 和 k，对于长度为 n 的输入数据，[263] 计算可以在至多 An^k 步内完成(对任意的 n 值).

例如将两个整数相加(手算)的标准算法是多项式时间算法. 如果整数用标准的十进位表示，基本运算是两个数字相加，那么将两个各有 $n/2$ 位数字(输入数据长度为 n)的整数相加恰好需要 n 步(包括进位)，因此满足上述定义且 A 和 k 皆等于1. 在两个各有 $n/2$ 位数字的整数乘法中，共有 $n^2/4$ 步基本数字乘法，加上 $n/2$ 步进位，

总计给出 $n^2/4+n/2$ 步. 因为 $n^2/4+n/2$ 永远小于 n^2,如果在上述定义中取 $A=1, k=2$,你就会明白整数乘法(按标准方法)也是一种多项式时间算法.

上述的例子如果是借助于图林机而不是十进数算术给出,你当然会用到常数 A 的更大的值,可能还需要更大的 k 值,但涉及的仍然是多项式时间算法. 事实上这就是为什么多项式时间算法的概念独立于机器与具体的程序,机器与程序的变换只会引起两个常数大小的改变,而定义本身依然有效.

不是多项式时间算法的算法被称之为指数时间算法. 例如一个算法处理长度 n 的输入数据时需要 2^n(或 3^n, n^n, $n!$)步,它就是一个指数时间算法. 这说明了这里用词"指数"的意义,虽然有可能引起误解,因为它包括了像 $n^{\log n}$ 这样的函数,而后者通常并不被看作为"指数"函数.

正如你目前已经了解的那样,"有效"算法是需要多项式时间的算法,而"非有效"算法则是需要指数时间的算法. 第1章中关于指数增长的讨论应当足以使你相信指数时间算法是高度的非有效算法(不过请看下文),而你大概会怀疑是否多项式时间算法必定是有效算法. 多项式时间定义中常数 A 和 k 选择的任意性似乎过于灵活:一个算法如果只有当你选择 $A=10^{10}$和 $k=100$时才能被分类为"有效"算法,那就很难被看作是真正意义上的有效算法. 这里有两点需要说明. 首先,实际情形是:问题最后要么只能用指数时间求解,要么则用有$10n^3$或更少步数的多项式时间算法来求解. 其次,多项式/指数时间的分法仅仅是一种粗略的、初步的分类. 将来可能需要寻找更为精细的区分,不过目前这种分法还很有用. 上述两点说明在表4中得到了强调. [264]

时间-复杂性函数	数据长度:n					
	10	20	30	40	50	60
n	0.00001s	0.00002s	0.00003s	0.00004s	0.00005s	0.00006s
n^2	0.0001s	0.0004s	0.0009s	0.0016s	0.0025s	0.0036s

续表

时间－复杂性函数	数据长度：n					
	10	20	30	40	50	60
n^3	0.001s	0.008s	0.027s	0.064s	0.125s	0.216s
2^n	0.001s	1.0s	17.9 分	12.7 天	35.7 年	366 世纪
3^n	0.059s	58 分	6.5 年	3855 世纪	2×10^8 世纪	1.3×10^{13} 世纪

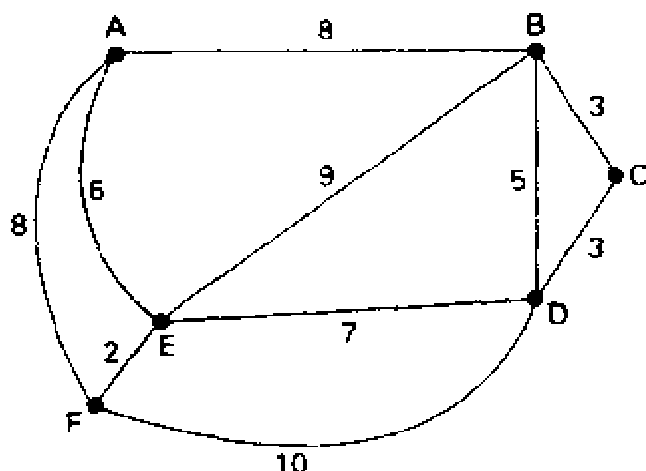


图63 旅行推销员问题.找出一条行遍所有城市的路线,使总的旅程达到最小.(图中标出了任意两地之间通路的距离.)例如旅线ABEFDC表示一次路程长为8+9+2+10+3=32的旅行.(有时要求旅行的起点与终点是同一地方,在此情况下那条特殊的路线可能是不完全的.)

机上实现,至少原则上是如此.但是即使对很少的城市来说,可能的路线数字也大得惊人.假设要访问 N 个地方,就会有 $N!$ 条可能的旅行路线.($N!$ 读作"N 阶乘",请回想一下,它是从 $N, N-1, N-2$ 到3,2,1的所有整数的乘积.)函数 $N!$ 当然属于指数型(它增长起来比 2^N 或 3^N 快,虽然比不上"超指数"函数 N^N),因此列出所有可能的路线显然将导致一个指数时间算法.为了明白这种方法情况有多糟,请注意如果有10个地方,就会有 $10! = 3628800$ 条可能的路线.这在现代计算机上还可以处理,但若有25个地方,那么可能的路线数就变成16后面再加25个零,这是一个吓人的数字.而在现实生活中一个推销员旅行25个城市并不是不可能的,更不用说涉及本质上是同类的数学问题的其他情形了,在那里"城市"数可能会成百上千. [266]

因此通过列举所有可能性来解决问题的做法显然是不可取的,除非涉及的城市数目很小.那么有什么其他办法可试吗?也许可以根据"常识"来解决问题?例如通过观察地图(或距离表),你可以确

定一条路线，首先走访出发点附近的城市，然后移向远处．虽然这样的方案（或者任何其他你愿意尝试的方案）对某些特殊情况可以行得通，但已经证明它并不是在所有情况下都行之有效的．我们现在所关心的是算法的总体行为，个别的推销员问题也许可以找到简单的解答（例如若所有的城市都位于同一条直线上，则最短路线是显而易见的），但我们要求的是在一切情形下都行之有效的算法．尽管自从1930年威尼斯数学家孟戈（Karl Menger）首次提出这个问题以来人们已经作了大量的努力（迄今已发表了300多篇这方面的专门论文），但寻找推销员问题一般解的所有尝试都告失败．事实上，正如我们将要看到的那样，有很充分的证据说明该问题根本就不存在有效算法！

然而在此之前，应当先介绍一下对某些特殊情况所取得的重要进展．首先是1962年，IBM公司的海尔德（Michael Held）和卡普（Richard Karp）用所谓动态规划的技巧对至多有13个城市的所有路线解决了推销员问题（13！ = 622702080）．1963年李特尔（Little）、墨梯（Murty）、斯维尼（Sweeney）和卡普发明了一种叫"分支限界法"（branch and bound）的强有力的技巧，他们靠这种方法在一台计算机（IBM 7090）上仅用了几分钟时间就解决了40个城市的推销员问题．1970年，海尔德和卡普又进一步发展了分支限界算法并用以解决推销员问题的另一个特例，即有42个城市的情形．这一算法只需要对数字庞大（33后接49个零）的可能路线检验其中的61条．（这个重要的特例涉及遍布美国的42座城市，1954年已被丹齐克（Dantzig）、福克斯顿（Fulkeston）和兰德公司的约翰逊（Johnson）解决．）1979年克劳德（Crowder）和派德伯格（Padberg）解决了有318座城市的特殊[267]情形，这在当时是曾经处理过的最大数字的特例了．这个问题目前的状况是：所用的方法应该能解决——在合理的计算机时间内，比方说几天内——城市数不超过500的任何特殊情形．然而个别特例自身的结构仍是关键的因素．大体上说，从现实生活中提出的问题——城市间的实际旅行——结果都能被证明是可解的，但人们可能设计出

将使一切现有的方法都无能为力的“人造”例子.在下一节中我们将要说明为什么对所有情况都适用的一般解法很可能并不存在.

P 和 NP

关于问题能否用有效算法求解的抽象讨论,终于使人们发现,如果把所有的问题都转化成只要求简单回答“是或非”的形式,将会带来许多方便,就可以对不同的问题进行比较.例如乘法问题(已知整数 a 和 b,它们的积等于多少?)可以转化为:已知整数 a,b 和 c,$ab=c$ 是否成立?推销员问题则可以转化为:已知一个城市的集合,以及这些城市间的距离表,对于给定的数 B,是否存在一条这些城市间的旅行路线,其总路程至多等于 B?(这问题与原来陈述的问题是否实质上相同,并不是显而易见的.但事实上它们是等价的.可以证明如果对原来的问题存在有效算法解,那么约化以后的问题也一定存在有效算法解,并且反之亦然.)

只要求回答“是或非”的问题叫做判定问题(decision problem).一个判定问题被称为 P 型问题,如果它可以用多项式时间算法求解.例如上述乘法问题就是 P 型问题.要判断 $ab=c$ 是否成立,只要将 a 和 b 相乘,然后看结果是否等于 c.这只需要多项式(实际上是二次式)时间.

旅行推销员问题可能是 P 型问题,也可能不是 P 型问题,迄今为止对此尚无定论.已经知道的是该问题是所谓 NP 型问题,NP 意思是“非确定的多项式时间”.为了理解这一概念,设想有一台图林机(或者其他任何计算机),它能在运算的各个步骤进行随机猜测.(这只能想象,因为不可能造出这样的计算机.)利用这样的假想计算机 [268] (称做非确定型图林机),推销员问题就能用多项式时间求解.算法很简单.首先猜测第一个要访问的城市,然后猜第二个,第三个,如此等等,直到猜完整个旅行路线.然后计算总的路程并与已知数 B 比较.只要机器每一步都“猜测正确”(现实中一个或然事件发生的概率为

$1/N!$，N 是待访问的城市数），最后所得的结果也就正确．这就是所谓 NP 型问题的涵义：通过一次或多次"正确的"（或"最优的"）猜测，问题可能在非确定型图林机上用多项式时间求解．

另一个 NP 型问题是检验整数是否为合数（即非素数）．为了检验一个给定的整数 n 是否是合数，首先猜测两个小于 n 的整数 a 和 b，然后验算是否有 $ab = n$．其答案显然可以在多项式时间内得到，而最优的猜测就是问题的正确答案．然而请注意，同样的方法不足以证明"补问题"即判断一个整数是否为素数的问题属于 NP 型．为了证明一个数 n 是合数，所需的一切就是一个好的猜测，而为了证明 n 是素数，一切猜测都无济于事．事实上素性检验是一个 NP 型问题，但要证明这一点，你必须使用完全不同的素性检验方法．

NP 型问题这一高度抽象的概念具有重要的意义，这是由两方面的原因共同决定的．首先是许多尚未找到有效算法的问题被证明是属于 NP 型．（我们可以直觉地看出，在这些问题中困难不是来自必须的计算过程，而是在于存在着数量巨大的可能性．因此当所有这些不同的可能性机会均等能以同样的方式来处理时，猜测对策就很适用，而这正是 NP 概念的基础．）这样 NP 就为大量实际问题的解决提供了理论框架．

第二方面的原因是由库克（Stephen Cook）1971 年关于算法有效性的工作引起的．借助于图林和其他人的方法，库克提出一种方法来证明某些 NP 型问题极不可能用有效的多项式时间算法求解．（正如我们将要看到的那样，这里的措辞"极不可能"．实际上可以用"肯定不能"来代替．）特别是库克证明了存在着一类特殊的 NP 型问题，他称之为 NP 完全性问题（NP – complete），其意思是：如果这类特殊的问题可以用多项式时间算法求解，那么所有其他 NP 型问题也都可以用多项式时间算法求解．换句话说，库克所考虑的这个问题与所有其他的 NP 型问题同样"硬"．利用库克的结果，后来有些数学家证明了还有许多其他的 NP 型问题也是 NP 完全性问题，其中包括旅行推销员问题（请看框图 C）．

[269]

框图 C:某些 NP 完全性问题

旅行推销员问题(详见正文)

哈密顿环路问题 已知一个由一些城市和连结这些城市的道路组成的网络,是否存在一条旅行路线,使起点和终点都在同一个城市,而其他每个城市恰好都经过一次?

多道加工机任务安排问题 给定一系列需要完成的任务集 T,以及在某类加工机上完成每件任务所需的时间表,同时还给定了这类加工机的具体台数,是否可能将 T 中的任务分组并将每组任务分配给一台加工机,使得完成所有这些任务的总时间小于某个指定的时间?(每台加工机顺序连续工作,而全体加工机则同时运行.)

地图着色问题(参见第 7 章) 已知一张地图,是否能够只用四种颜色来着色,使任何两个有公共边界的国家着色不同?

二次剩余问题 已知正整数 a,b,c,并有 $a<b$,是否存在一个正整数 $x<c$,使 $x^2 \bmod b = a$?

二次丢番图方程(参见第 6 章) 已知正整数 a,b,c,是否存在正整数 x 和 y,使 $ax^2+by=c$?

[270]

于是,作为库克等人工作的结果,人们得以证明从现实世界提出的许多(NP 型)问题与任何其他的 NP 型问题一样地难.大多数数学家现在会认为:既然已经知道一个问题与任何其他 NP 型问题一样难,再去寻找解决这问题的有效(亦即多项式时间)算法就是枉费时间了.因此难免会产生这样的看法:如果证明了一个给定的问题是 NP 完全性问题,就等于证明了它不能用多项式时间算法求解.

然而困难在于:库克的结果(以及所有以后得到的结果)并没有排除这样的可能性,即类 P 与类 NP 实际上是相同的,也就是说任一 NP 型问题实际上都可用多项式时间算法求解(虽然在特殊的例子中要找到这样的算法并非易事).如果情况真是如此,那么知道一个问题"像其他任何 NP 型问题一样难"实际上毫无意义.(在这样的情况

下所有的 NP 型问题都一样“容易”.)不过很少有专家认为存在这样的可能性.NP 概念本质上涉及非算法的“猜测”过程(确切地说是“猜对”,以区别于通常的碰运气),这就使得它不可能完全等价于 P 型问题.因此,P 和 NP 在理论上可能等价的说法从一开始就大受怀疑,而 NP 完全性才被看作是问题真正“不可解”的证明.

当然,全部问题的解决其实只需要一个反例,即找出一个属于 NP 型但却能证明不是 P 型的问题.然而尽管 P 和 NP 这两个概念直觉有别,问题却还远没有得到解决,并且各方面的迹象都表明这是一个极其困难的问题.这个以“P = NP 问题”著称的问题,正如你可以想象的那样,已经成为当今计算机数学中最重要的未决问题之一.这问题之所以重要,部分原因当然是在于它与大量实际问题有关,但这里我们务需谨慎,因为有关的问题都不简单,不能等闲视之.因此,现在 [271] 是我们回到现实生活的时候了.

回到现实生活——线性规划

虽然刚才介绍的理论方法可以提供一些有价值的信息,但他们对于应用问题却并不是总能给出精确的描述.一个在理论上需要指数时间的算法(即“非有效”算法),在实践中处理日常数据时却可能用得很好.指数运算时间可能仅仅是产生于某些不常遇到的数据.推销员问题在某种意义上就属于这一范畴.对“现实生活”中的城市与道路布局,已有的方法——它们无疑都需要指数运算时间——应用得很好.关于理论与实际之间这种潜在的鸿沟,所谓的线性规划问题提供了更为惊人的例子.线性规划问题孕育了一门叫运筹学的学科,并且至今仍是这门学科的中心课题.(运筹学产生于第二次世界大战期间,是借助于数学方法来解决与工业、商业、政府、国防等部门人力、机械、物资、钱财等大系统的指导、管理有关的复杂问题的学科.)

有一类实际问题需要将某些对象最大化(如利润、安全等)或最小化(如支出、风险等),线性规划就是为这类实际问题提供数学描述

(或模型)的一种方法.所要求的最优化(optimization)是通过适当选定一些参数(或变量)的值来实现的.无论是待优化的因素,还是部分或全部的参数,都必须服从于一个或多个约束条件.线性规划中"线性"二字是表示模型中所有的数学表达式都是线性的(即不涉及两个或多个变量的乘积或一个变量的乘幂).这在实践中不能算是太大的限制,因为大多数实际问题或者本来就是线性的,或者可以被假定是线性而不会产生任何大的误差.

对这个问题的初步考察就可以看出,线性约束有一个很自然的几何表示.满足所有约束的变量的值相当于某个几何图形内部的点.如果有两个变量,这一图形就是一个多边形(边数对应于约束数),如果有三个变量,这一图形就是一个多面体,而如果有 N 个变量,这一图形就是 N 维空间中的一个多胞形(参见第10章).当然我们不可能画出四维或更高维空间中的多胞形,但不论是几维空间,这里用到的数学是简单的. [272]

可以用一个简单的例子来说明这一切.设有一家公司生产两种布料 A 和 B,所用原料是三种不同颜色的羊毛.表5给出了生产单位长每种布料所需的羊毛量,以及可供使用的每种颜色的羊毛的总量.生产者的利润是:A 布每单位长£12,B 布每单位长£8.问题是:怎样使用所供的羊毛使可能获得的总利润最大?

羊毛颜色	单位长需要量		供应量
	布 A	布 B	
红	4kg	4kg	1400kg
绿	6kg	3kg	1800kg
黄	2kg	6kg	1800kg

表5 生产单位长 A 布与 B 布所需的红、绿、黄三种羊毛量,以及可供使用的羊毛总量.

我们首先设 x 和 y 分别表示生产出的 A 布量和 B 布量. 这样就将产生利润 P(£)

$$P = 12x + 8y. \tag{10}$$

x 和 y 值所受的约束是什么呢? 因为只有 1400kg 红羊毛可供使用, 而生产两种布每单位长所需的红羊毛都是 4kg, 所以

$$4x + 4y \leqslant 1400. \tag{11}$$

类似地考虑可供使用的绿色和黄色羊毛量就得到

[273]

$$6x + 3y \leqslant 1800, \quad 2x + 6y \leqslant 1800. \tag{12}$$

最后, 因为无论 x 或 y 都不应当是负数(这个约束在所讨论的实际问题中是显然的, 但在数学表述中需作明确说明), 故需有

$$x \geqslant 0, \quad y \geqslant 0. \tag{13}$$

图 64 就是由不等式(11)、(12)和(13)给出约束的图示. 满足所有约束的任何一对 x 和 y 的值都将是图中阴影区域内一点的坐标; 反之该区域内任何一点的坐标都将满足不等式(11)、(12)和(13). (请你自己通过读出该区域内部或外部各点的坐标来检验上述结论.) 于是我们现在要做的事情就是在带阴影的区域内找出一点, 使方程(10)中的量 P 尽可能大.

[274]

现在来考虑方程(10)所表示的直线, 对于任意确定的 P, 方程(10)所表示的直线都是互相平行的. (图 64 中对 $P = 1200$ 和 $P = 2400$ 给出了两条这样的直线.) 因此十分清楚, 为了使 P 最大化, 我们必须这样做: 平行移动(由方程(10)给出的)利润直线, 使它尽可能地远离原点但又不完全脱离阴影区域. 这就把我们带到了 B 点. B 点的坐标很容易通过初等代数而确定(两个联列方程的解): 它们是(250, 100). 因此, 为了获得可能的最大利润 £3800, 厂商必须生产 A 布 250(单位), B 布 100(单位). (这刚好用完了所有可供使用的红色的和绿色的羊毛, 但却剩下了 700kg 黄色羊毛. 我们的厂家在购进羊毛以前要是能做一次数学家庭作业就好了.)

在解决了这个具体问题以后, 我们来考察一下它说明些什么. 约束条件在图 64 中被表示成一个平面多边形区域 ABCDO. 最大化点

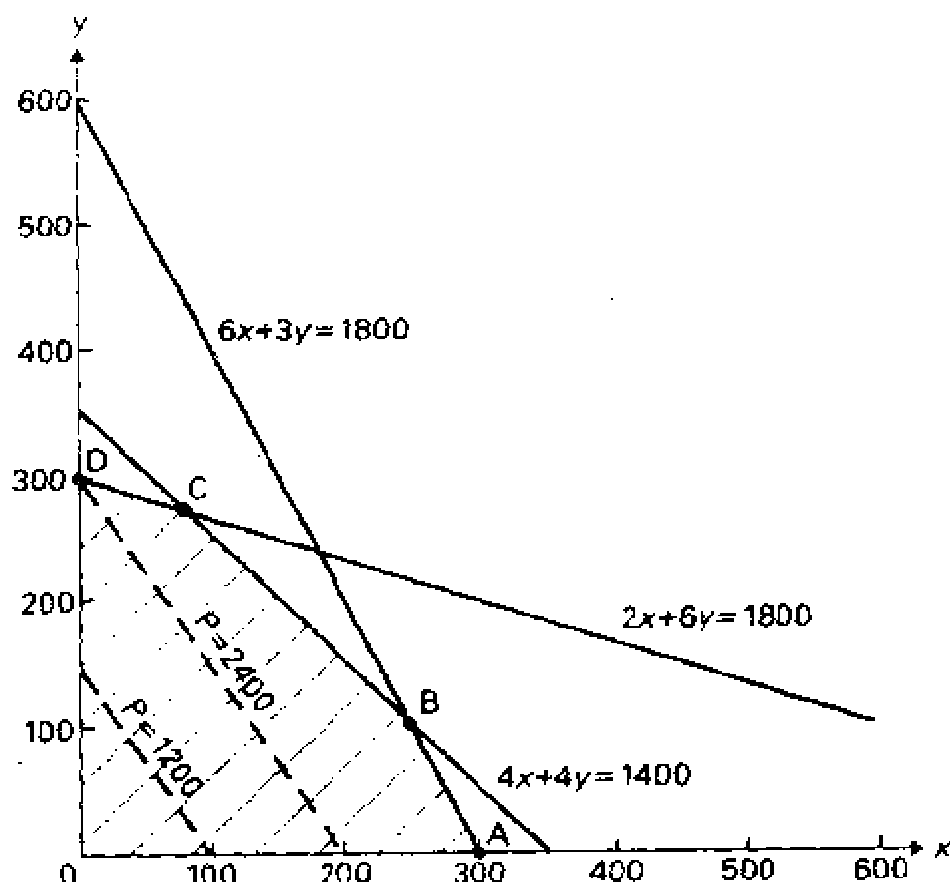


图 64 线性规划.布料生产问题的解[1](详见正文).

是这个多边形的顶点——剩下的任务就是找出这个顶点.在上述简单例子中,找出这样的顶点并不困难,但对于更复杂(因此也更实际)的线性规划问题来说,这正是难点所在.在有三个变量的问题中,约束条件将产生一个三维的多面体;而在N变量的情形你将会得到一个N维的多胞形.这样的多胞形不可能被画出来,但仍能进行代数的处理.在每个场合,问题都被归结为寻找约束图形(多边形,多面体或多胞形)的最优化顶点.但是怎样才能做到这一点呢? 可能会有上百万个(甚至更多)顶点! 正如像推销员问题的情形一样,这时用穷举的方法是解决不了问题的.因此就需要有确定最优顶点的系统方

① 原书本图阴影区域标示有误,译者已作了更正.——译者注.

法.

1947 年,美国数学家丹齐克(George Dantzig)设计出了一种这样的方法:单纯形算法(simplex algorithm).这种方法实质上就是从一个顶点出发(怎样准确地确定这个起始点在这里略去不谈),然后在多胞形表面沿着边从一个顶点向另一个顶点移动.每达到一个新的顶点,就将有几种(或很多种)不同的继续移动路线,并且有各种判别方法来决定究竟走哪条路线(最"显然"的决定方法就是只向使需最大化的量有所增加的顶点移动,或在最小化情形只移向使最小化量有所减少的顶点).

由于沿着一个多胞形的边进行的可能的路线数量很大,单纯形算法在理论上被认为是一种指数时间算法,然而在实际应用中(对于[275]涉及几百甚至几千个变量的问题)这算法却非常有效,可以在相对而言较少的步骤内找到最优顶点.事实上许多迹象都表明它可以在线性时间内完成运算(即运算步数与涉及的变量个数成正比).导致该算法非有效运作的那种约束(以及相应的多胞形)在实际问题中并不常见——它们必须特别"泡制"出来,而目的只是为了使单纯形法败阵而去.但事实是无论怎样精心策划,这些"人造"的问题却丝毫不能阻挡单纯形法的发展:每当有一种新的工业或商用计算机系统投放市场,应运而生的第一批商品中总少不了修订版的单纯形法!一句话,这种方法行之有效.

那么,难道就没有更快的方法了吗?一种不仅对大多数问题而且在所有的情形下都更快的算法——也就是多项式时间算法?从直观上看,如果不是在多胞形表面沿着边前进而是直接穿过多胞形内部去寻找最优顶点,照例说应该更快些.但这种做法的困难在于:因为你事先并不知道哪一个是最优顶点,又怎样来决定前进的方向呢?如果是停留在多胞形表面,你至少还有一种办法来决定每一步怎么走,但是一旦你"脱离"表面而走进多胞形体内,那么究竟有没有方法为你导向呢?

事实上是有的.1970 年,苏联数学家肖尔(Shor)发现一种叫牛顿

法的老方法可以被应用于线性规划问题.这一想法后经莱文(Levin)、朱丁(Judin)和纳米洛夫斯基(Nemirovski)(也都是苏联人)等改进并于1976年发展为所谓椭球法,在这种方法中,穿过多胞形内部的路径其方向是借助于一系列被画出来"逼近"多胞形的椭球来决定的.1979年喀奇安(Khachian)(也是苏联人,他对上述技巧运用自如)证明了椭球法是多项式时间算法.不幸的是,虽然这意味着这种方法在理论上比单纯形法强,但在实用中却并没有显示出任何比单纯形法优越的地方.

因此你可能会说:"既然理论上非有效的方法可以轻而易举地胜过理论上有效的方法,那就让理论家的有效性概念见鬼去吧!"许多非数学家也确实说过类似的话.毕竟线性规划问题曾经是并且现在仍然是最重要的实际问题.如果有什么例子揭了多项式时间/指数时间有效性分类法的短,那么从纯粹数学家的观点看,线性规划问题就是最糟糕的一个.

[276]

但是接下去,在1984年初,另一位理论家登场了.卡玛卡(Narendra Karmarkar),一位在美国贝尔实验室工作的28岁的数学家,发现了一种多项式时间的线性规划算法,这种算法不仅在实用中确实行之有效,而且在许多场合竟远远超过了单纯形法.(在一次对有5000个变量的问题的试验中,卡玛卡算法比单纯形法要快50倍.)这是一个引人注目同时又出人意料的进展,卡玛卡为了得到他的新算法,不得不使用了某些高度成熟的数学,包括一系列多胞形"整形"(reshapings)以确定在其内部前进的"优先"方向.(虽然,正如单纯形算法在计算机上使用时需要修订,因为计算机只能进行算术运算而不能直接处理背后的几何概念,卡玛卡算法的修订版也抹去了其中成熟的几何概念以便于进行矩阵的算术运算.)

因此,理论家的有效性概念最终被证明是行之有效的.而且,新的算法提供了一个惊人的例子,说明一些高度成熟的抽象数学,其中包括多面体的高维类似和稀奇古怪的数学"变形",怎样会导致对商业和国防等现实世界有重大意义的具体成果.这可以说是纯粹、抽象

的数学与我们生活的现实世界相结合的典范，我们对数学的新黄金时代的介绍，最好就以此结束.

阅 读 文 献

关于算法有效性的可读性很强的简要说明可见 H.Lewis 和 C.Papadimitriou 的文章 The efficiency of algorithms，刊于 Scientific American，Volume 238(January 1978)，pp.96 – 109. 学习算法有效性的标准的入门教科书是 M.Garey 和 D. Johnson 的 Computers and Intractability (W.H.Freeman，1979).

关于线性规划问题和单纯形算法的初等介绍，可参阅 J.Lighthill 编的 New-[277] er Uses of Mathematics(Penguin，1978).

关于 Karmarkar 的算法的介绍，可参阅他本人的文章 A new polynomial-time algorithm for linear programming，刊于数学杂志 Combinatorica，Volume 4(1984)，[278] Number 4，pp.373 – 92.

(李文林译)

人名索引

专名索引